AD_______________

AWARD NUMBER DAMD17-97-1-7130

TITLE: Computer-Assisted Visual Search/Decision Aids as a Training Tool for Mammography

PRINCIPAL INVESTIGATOR: Calvin Nodine, Ph.D.

CONTRACTING ORGANIZATION: University of Pennsylvania
Philadelphia, Pennsylvania 19104-3246

REPORT DATE: July 1999

TYPE OF REPORT: Annual

PREPARED FOR:
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for public release; distribution unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
July 1999

3. REPORT TYPE AND DATES COVERED
Annual (1 Jul 98 - 30 Jun 99)

4. TITLE AND SUBTITLE
Computer-Assisted Visual Search/Decision Aids as a Training Tool for Mammography

5. FUNDING NUMBERS
DAMD17-97-1-7130

6. AUTHOR(S)
Nodine, Calvin, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Pennsylvania
Philadelphia, Pennsylvania 19104-3246

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research And Materiel Command
ATTN: MCMR-RMI-S
504 Scott Street
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The primary goal of the project is to develop a computer-assisted visual search (CAVS) mammography training tool that will improve the perceptual and cognitive skills of trainees leading to mammographic expertise. In two years we have completed two studies. The first equates experience by comparing perceptual skills of expert radiologists with laypeople searching non-medical pictorial scenes for hidden targets. Results show that expert radiology search and detection strategies do not transfer to the non-medical search and detection tasks. Thus, radiology expertise consists of specific perceptual and cognitive skills that develop primarily from experience reading medical x-ray images. The second study examines the roles of training and experience on the acquisition of mammography expertise. We compared reading performance of experienced mammographers with residents at different levels of training and technologists with little training or experience reading mammograms. Results show, first that performance is a linear increasing function of log reading experience, second that mammography training provides insufficient reading experience to meet clinical standards of performance, third that expert mammography performance is characterized by a speed-accuracy relationship that is the result of tuning of visual search and recognition skills acquired primariy through practice reading mammograms with feedback.

14. SUBJECT TERMS
Breast Cancer

15. NIIMRER OF PAGES
68

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
Unclassified

18. SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

19. SECURITY CLASSIFICATION OF ABSTRACT
Unclassified

20. LIMITATION OF ABSTRACT
Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

N/A Where copyrighted material is quoted, permission has been obtained to use such material.

N/A Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

N/A Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

N/A In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

CFN For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

N/A In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

N/A In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

N/A In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

CFNodine 7/16/99

PI - Signature Date

TABLE OF CONTENTS (DAMD17-97-1-7130,Revised 7/31/99)

(1) Front Cover

(2) Standard Form (SF) 298, Report Documentation Page (SF-298)

(3) Foreword

(4) Table of Contents

(5) Introduction

(6) Body

6.1 Objectives

6.2 Technical Objective 2,Task 3: Collect Mammogram Cases; Digitize Collected Mammograms; Construct a Training Set of Breast Lesion Images.

6.3 Technical Objective 2, Task 4: Obtain MAMCAD; Integrate Training Set Into MAMCAD.

6.4 Technical Objective 1, Task 1: Program ASL MODEL 4000; Task 2: Modify Eye-Position Data Collection Programs and Develop and Integrate Detection Algorighm Within PC WINDOWS 95 Environment.

6.5 Completion of the Resident Study (Expertise Study).

6.6 Summary of Mammography Expertise Study.

6.7 Implications of Expertise Study.

(7) Key Research Accomplishments

(8) Reportable Outcomes

(9) Conclusions

(10) References

(11) Appendices

PROGRESS REPORT, 1998-1999, Year 2, DAMD17-97-1-7130 COMPUTER-ASSISTED VISUAL SEARCH/DECISION AIDS AS A TRAINING TOOL FOR MAMMOGRAPHY.

C.F. NODINE, PI

(5) INTRODUCTION:

This project focuses on the training of diagnostic interpretation skills in mammography which are acquired primarily as a result of medical training and experience reading mammograms. These skills take years of formal training and mentoring experience with experts who help interpret and illustrate a variety of abnormalities in breast images. Although medical training is typically rigorous and systematic, the mentoring experience during radiology residency in mammography varies widely from one teaching institution to another. The primary aim of this project is to develop a computer-assisted mammography training tool that will act as a surrogate mentor in aiding radiology residents in making plausible diagnostic decisions. We emphasize plausible diagnostic decisions as the end result of medical problem solving, because in clinical mammography resident rotations, diagnostic truth is typically defined as agreement after mammographic assessment between a radiology resident and mentor rather than by gold-standard pathologic truth. We propose to provide computer aids that will interact with the resident immediately after image interpretation by providing systematic feedback about visual search, detection and decision making. This feedback will point out, by highlighting, what areas of the mammogram receive prolonged visual dwell and decision time. These two parameters, visual dwell and decision time, predict regions of suspicion on the mammogram (Krupinski, Nodine, Kundel, 1998; Nodine, Kundel, Mello-Thoms et al., 1999). The resident is then asked to reexamine the highlighted areas, determine if any abnormal features are present, and reevaluate the original diagnostic decision. This reevaluation of suspicious regions with feedback provides the basis for a plausible problem-solving solution based on the individual observer's initial perceptual analysis of the mammogram. We showed in 1990 (Kundel, Nodine, Krupinski, 1990) that computer-assisted visual search (CAVS) is effective in improving the detection of lung nodules, and Krupinski (1996) showed that visual dwell predicts the location of true and false, positive and negative decision outcomes. Our goal is to apply CAVS to mammography training to see if we can enrich the learning experience of radiology residents during training and thus improve their diagnostic interpretation and problem-solving skills.

(6) BODY:

(6.1) OBJECTIVES. Work Completed from July 1, 1998 to June 31, 1999 based on the approved Statement of Work.

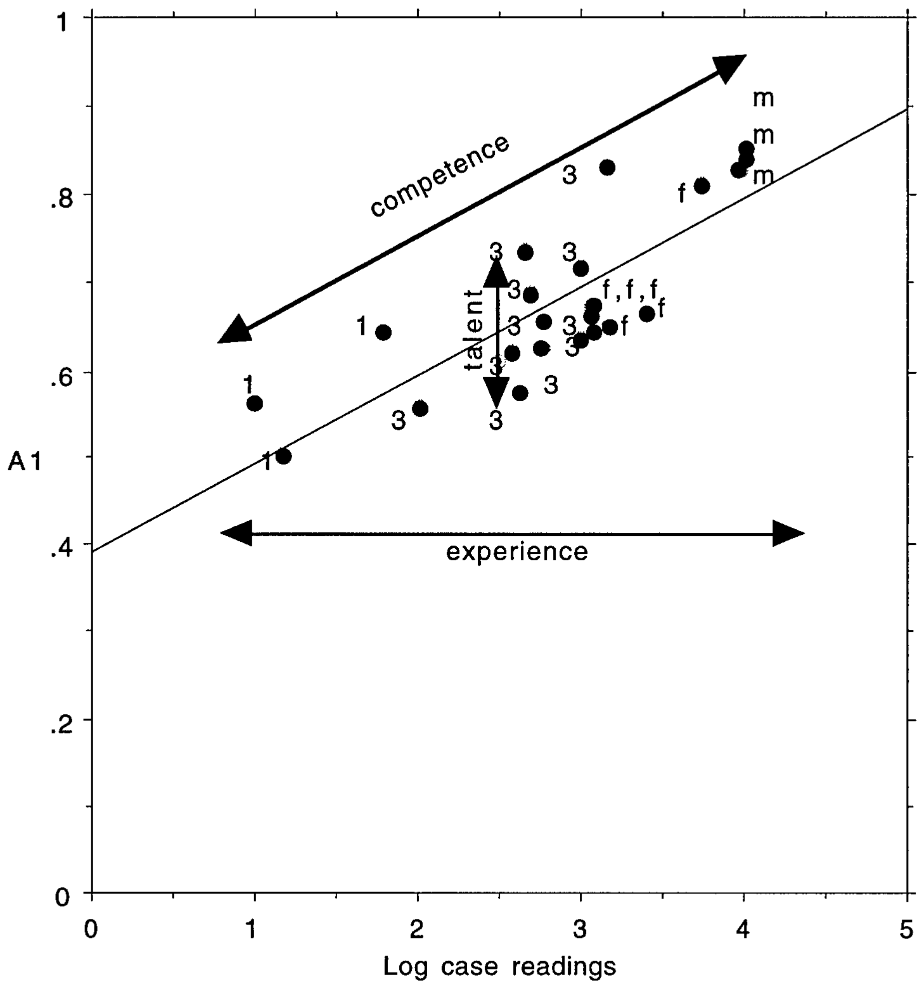
(6.2) TECHNICAL OBJECTIVE 2, TASK 3: COLLECT MAMMOGRAM CASES; DIGITIZE COLLECTED MAMMOGRAMS; CONSTRUCT A TRAINING SET OF BREAST LESION IMAGES. As stated in the Progress Report for year 1, we have been working on Technical Objective 2. We have completed digitization and construction of the training set of mammograms and developed three test sets of mammograms from it. One test set consisting of 75 mammogram cases (2 views, cc and mlo) was used in the expertise study (See Appendix 1). A second test set consisting of 40 mammogram cases is currently being used in a study designed to measure the role of eye fixations in visual search and detection of subtle lesions. In total, we have digitized 325 mammogram cases (1150 images) of which all but 75 have been digitized at 50 micron pixel size. The 50 micron cases will be used to construct a training set of mammograms to be used in the CAVS study in years 3 and 4. **TASK 3 COMPLETE.**

(6.3) TECHNICAL OBJECTIVE 2, TASK 4: OBTAIN MAMCAD; INTEGRATE TRAINING SET INTO MAMCAD. Instead of obtaining MAMCAD from Dr. Alastair Gale at the University of Derby, we have decided to develop our own alternative to MAMCAD for Task 4 which, rather than analyzing the image for features signaling abnormality, analyzes the observer's decision time and confidence for correctness based on predictions derived from ANN. This ANN program, as we presently conceive it, would prompt the observer about the probability of making a false positive after decision time exceeded the observer's threshold for true positives. We have developed an Artificial Neural Net (ANN) program based on decision time data obtained from the expertise study (See Appendix 1). This program uses decision time to differentiate true positives from false positive decision outcomes for individual observers having different levels of mammography interpretation skill. We are currently testing the efficacy of this ANN program as part of TASK 5. Finally, as part of TASK 5, we have completed the Resident Study which is now referred to as the "Expertise Study" (See Appendix 1). the main findings are summarized below. **TASK 4 COMPLETE; TASK 5 IN PROGRESS.**

(6.4) TECHNICAL OBJECTIVE 1, TASK 1: PROGRAM ASL MODEL 4000; TASK 2: MODIFY EYE-POSITION DATA COLLECTION PROGRAMS AND DEVELOP AND INTEGRATE DETECTION ALGORITHM WITHIN PC WINDOWS 95 ENVIRONMENT. We have completed TASK 1. We have programmed the ASL Model 4000 to monitor the observer's eye position relative to head motion for digital mammography displays. We have also completed TASK 2. We have modified eye-position data collection programs (EYEPOS/EYEDAT) to accommodate a visual-dwell detection algorithm, and tested the integration of the detection algorithm for use with visual feedback of dwell locations on the PC display workstation. It will be necessary to send eye position data from the ASL 4000 computer to the Windows 95 environment, analyze eye fixations and identify image locations that receive prolonged visual dwell. This has created a problem for us because the Windows 95 requires a special software driver (DLL) to communicate between the ASL 4000 DOS system and Windows 95. We do not have the technical expertise to write a DLL Windows 95 software driver, but have consulted with ASL and they have agreed to solve this problem for us. It will cost $3500. which is available in the equipment budget. **TASK 1 COMPLETE; TASK 2 IN PROGRESS.**

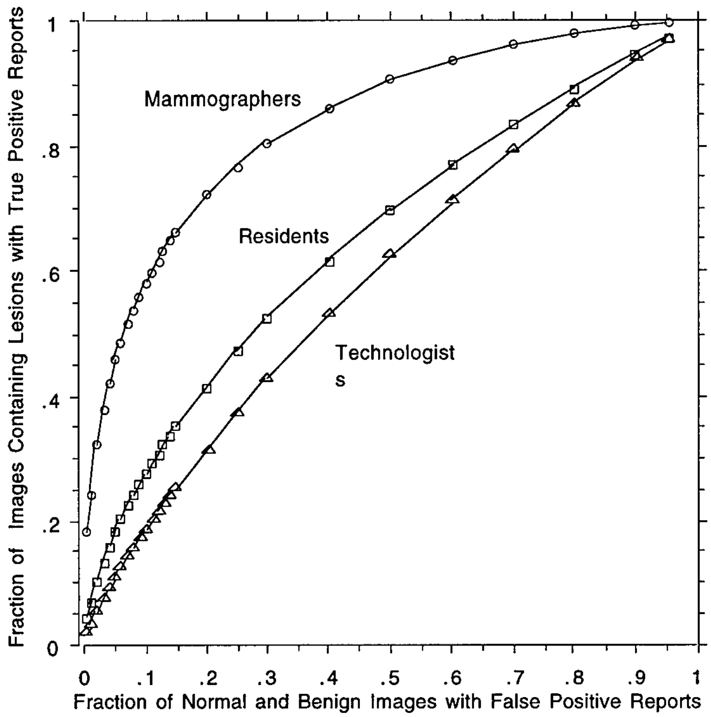
(6.5) COMPLETION OF THE RESIDENT STUDY(EXPERTISE STUDY). In order to complete Task 5, we carried out a study that examined how training and experience influence mammography expertise using a subset of digital mammograms from the breast-lesion image training set developed in Task 3. This was known as the "Resident Study" and is now referred to as the "Expertise Study". The purpose of this study was to explore how training via clinical mammography rotation influences several aspects of perceptual performance in breast screening by comparing three levels of mammography expertise exemplified by mammographers, radiology residents and mammography technologists who differ in both training and experience reading mammograms. The research question was: How does training affect residents' accuracy in visually differentiating malignant from benign lesions in a simulated mammography screening task? The answers are summarized below.

(6.6) SUMMARY OF MAMMOGRAPHY EXPERTISE STUDY. The complete final draft of this paper can be found in Appendix 1. The key findings from this study can be summarized as follows. Figure 1 shows the result of a regression analysis of overall detection performance measured as A1, the area under the Alternative Free Receiver Operating Characteristic curve, as a function of log (base 10) number of cases read over a 3-year period by 3 experienced mammographers and 19 radiology residents and fellows undergoing clinical mammography residency rotation. Figure 1 shows a significant linear-regression fit of the data (R^2= .667) with a positive slope suggesting that reading-skill as reflected by A1 performance increases directly with log case-reading experience (F (1,22)= 44.15, p<.0001).

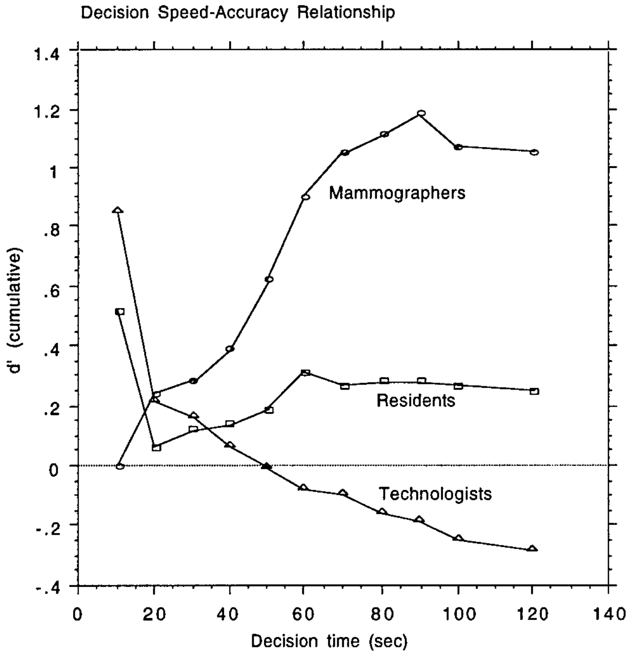


When case reading experience is zero, the regression line intercepts the y-axis at A1=.393 which is close to chance performance (0= chance). With mentor-guided case-reading training and experience, A1 performance increases. The numbers and letters within Figure 1 indicate the level of training of the observers: 1 = first- and second-year residents, 3 = third- and fourth-year residents, f = fellows and m= mammographers. Overall performance, A1, increases as a function of log (base 10) case reading experience. Figure 1 indicates that overall performance is directly related to training and experience reading mammograms with mentor guidance.

As a result of lack of reading experience, radiology residents performed significantly below their radiology mentors (Average A1= .840 for mentors vs. Average A1= .653 for residents, p<.01, Scheffe test). Residents performed at about the same level as mammography technologists (A1= .592 for technologists). This is shown in Figure 2.



The lower performance of radiology residents and mammography technologists is due to both failure to recognize true lesions (misses) and errors of commission (false positives). These differences in mammography reading skill are reflected in the speed-accuracy relationship shown in Figure 3.



The speed-accuracy relationship is measured by d', the index of detectability, as a function of decision time for mammographers, residents and technologists. Overall performance as measured by d', which is the normal deviate, z, of true positives/false positives, increased for mammographers and to a lesser extent for residents. Overall performance decreased below chance (d'= 0) for technologists meaning that false positives actually outnumbered true positives.

(6.7) IMPLICATIONS OF EXPERTISE STUDY. The analyses of decision times in the Expertise Study reveal significant differences in perceptual discrimination, object-recognition and decision-making skills among three levels of expertise, which shed light on the need for systematic feedback training during the radiology residency experience. The impact of this type of training on the mammography expertise of radiology residents will be evaluated in a formal experiment after the development of CAVS. Future studies will explore the use of computer-assisted visual search (CAVS) as a training tool that provides systematic visual feedback and decision aids to improve residents' detection and classification of distinctive pathologic features that differentiate malignant from benign breast lesions.

(7) KEY RESEARCH ACCOMPLISHMENTS:

Our research studies have led to three key findings:

1. Overall mammography diagnostic performance is log linearly related (base 10) to mammography reading experience according to a power law. This implies that reading experience is a major independent variable in the acquisition of mammography expertise.

2. A key characteristic of mammography expertise is expressed by a speed-accuracy relationship.

3. Mammography reading experience tunes visual perception and recognition skills underlying the speed-accuracy relationship.

(8) REPORTABLE OUTCOMES:

In addition to the work completed and in progress as discussed above, we have completed the following articles:

1. "How Experience and Training Influence Mammography Expertise", ACADEMIC RADIOLOGY, 1999, in press (see Nodine, Kundel, Mello-Thoms et al., 1999, in press, Appendix 1).
2. "A Chronometric Analysis of Mammography Expertise" was presented at Medical Imaging 1999, SPIE PROCEEDINGS 1999; 3663:146-150 (see Mello-Thoms, Nodine, Kundel, 1999, Appendix 2).
3. "A Perceptually Tempered Display for Digital Mammograms", was presented at RSNA, 1998, RADIOGRAPHICS, 1999, in press (see Kundel, Weinstein, Conant, Toto, Nodine, 1999, in press, Appendix 3).
4. "Enhancing Recognition of Lesions in Radiographic Images Using Perceptual Feedback" OPTICAL ENGINEERING 1998;37:813-818 (see Krupinski, Nodine, Kundel, 1998, Appendix 4).

We are currently working on two papers:

1. "The Effects of the First Response in Mammogram Reading: The Noise in the Head" which was presented at the Far West Image Perception meeting in British Columbia by Claudia Mello-Thoms.

2. "The Nature of Expertise in Radiology" by CF Nodine & C Mello-Thoms which will be Chapter 8 in the SPIE Medical Imaging Handbook in 2000.

(9) CONCLUSIONS

The primary goal of the project is to develop a mammography training tool that will improve perceptual and cognitive skills of observers leading to mammographic expertise. Prerequisites to this goal are an understanding of: (a) how mammographers are trained, (b) what skills are required to carry out the task of detecting, classifying and diagnosing abnormalities in mammograms, and (c) the effectiveness of current mammography training measured by evaluating the performance of residents using a test-set of mammograms representing various abnormalities.

We are beginning to understand how mammographers are trained. Training consists primarily of an apprenticeship with an expert mammographer who serves in a mentoring relationship with radiology residents. The mentoring relationship is carried out during radiology residency in a series of clinical rotations consisting of two 4-week experiences. During the clinical rotations residents read both screening and diagnostic mammographs with a mentor. Residents compare diagnosis with mentors and receive feedback not only about the correctness of their decisions, but more importantly, the reasons behind these decisions. In reading mammogram cases, they learn to use BI-RADS to categorize and report diagnostic decisions. Part of their residency rotation experience also consists of follow-up procedures after diagnosing breast abnormalities. We have show that the amount of experience reading mammogram cases with a mentor (defined as deliberate practice) has significant impact on overall diagnostic performance. The residents that we studied at the University of Pennsylvania received an average of 645 case-reading experiences which from our regression analysis leads to a performance prediction that is well below acceptable clinical standards (see Figure 1). This brings us to the question of what skills need to be improved, and how can this be accomplished.

Our research has focused on perceptual and decision-making skills in mammography. We have used eye-position recording to shed light on the role of visual search in diagnostic performance. Visual search skills translate into rapid image-perception assessment which leads to fast, accurate decision making as indicated by decision-time analyses. We have called this the speed-accuracy relationship (see Figure 3). Case-reading experience plays a key role in the speed accuracy relationship as shown by the shape of the d' curves which reflects overall performance (roughly true positives-false positives) as a function of decision time for different levels of mammography expertise.

Finally, when we come to the question of how can perceptual and decision-making skills be improved? The answer that our research seems to be saying is: "Practice Makes Perfect". This is a deceptively simple answer. During their medical training, radiology residents have to learn much more than how to read mammograms, and there is simply not enough time in the radiology residency program to make expert mammographers. Rather, what may be needed is a more effective way to train residents during their clinical residency in mammography. Maybe we need to supplement apprenticeship mentoring by expert computer systems. Expert computer systems can provide systematic feedback tailored specifically to each resident's level of training and experience. We propose to use CAVS, which can be "tuned" to provide systematic feedback about regions of the mammogram deemed "suspicious" based on analysis of eye-position dwell times. Prolonged visual dwells will be used to localize image regions for re-evaluation and decision making. Thus, CAVS may hold the key to more effective mammography training.

(10) REFERENCES

1. Beam CA, Layde PM, Sullivan DC. Variability in the interprtetation of screening mammograms by US radiologists. Arch Inter Med 1996; 156: 209-213.

2. Krupinski EA, Nodine CF, Kundel HL. Enhancing recognition of lesions in radiographic images using perceptual feedback. Optical Engineering 1998;37:813-818.

3. Krupinski EA. Visual scanning patterns of radiologists searching mammograms. Acad Radiol 1996; 3: 137-144.

4. Kundel HL, Nodine CF, Krupinski EA. Computer-diaplayed eye position as a visual aid to pulmonary nodule interpretation. Invest Radiol 1990; 25: 890-896.

5. Kundel HL, Weinstein SP, Conant EF, Toto LC, Nodine CF. A perceptually tempered display for digital mammograms. Radiographics, 1999, in press.

6. Mello-Thoms C, Nodine CF, Kundel HL. A chronometric analysis of mammography expertise. SPIE Conf on Image Perception and Performance 1999; 3663:146-150.

7. Nodine CF, Kundel HL, Mello-Thoms C, Weinstein SP, Orel SG, Sullivan DC, Conant EF. How experience and training influence mammography expertise. Acad Radiol, 1999, in press.

(11) APPENDICES

Appendix 1: Nodine CF, Kundel HL, Mello-Thoms C, Weinstein SP, Orel SG, Sullivan DC, Conant EF. How experience and training influence mammography expertise. Acad Radiol, 1999, in press.

Appendix 2: Mello-Thoms C, Nodine CF, Kundel HL. A chronometric analysis of mammography expertise. SPIE Conf on Image Perception and Performance 1999; 3663:146-150.

Appendix 3: Kundel HL, Weinstein SP, Conant EF, Toto LC, Nodine CF. A perceptually tempered display for digital mammograms. Radiographics, 1999, in press.

How Experience and Training Influence Mammography Expertise.

Appendix 1

Calvin F. Nodine, Ph.D. *, Harold L. Kundel, M.D., Claudia Mello-Thoms, M.S.E.E., Susan P. Weinstein, M.D., Susan G. Orel, M.D., Daniel C. Sullivan, M.D. †, Emily F. Conant, M.D.

University of Pennsylvania School of Medicine, National Cancer Institute†

*Corresponding Author
308 Stemmler Hall, 36th & Hamilton Walk
Department of Radiology, University of Pennsylvania Health System
Philadelphia, PA 19104-6086
215-662-6630
215-349-5115
nodine@oasis.rad.upenn.edu

This research was supported, in part, by Grant DAMD17-97-1-7103 between the USAMRMC and the first author.

Reprint requests should be addressed to:
Calvin F. Nodine, Ph.D.
308 Stemmler Hall, 36th & Hamilton Walk
Department of Radiology
University of Pennsylvania Health System
Philadelphia, PA 19104-6086
215-662-6630
215-349-5115
nodine@oasis.rad.upenn.edu

Running Title: Training and Mammography Expertise

ABSTRACT

Rationale and Objectives: Mammography expertise is characterized by fast, accurate diagnostic decision making. To evaluate the influence of perceptual and cognitive skills in mammography detection and interpretation, three groups representing different levels of mammography expertise in terms of experience, training and talent were tested using a combination mammography screening/diagnostic task.

Materials and Methods: A set of 150 mammograms, unilateral CC and MLO views, one-third malignant lesions, two-thirds malignant free, were displayed in pairs on a digital workstation to 19 radiology residents, 3 experienced mammographers and 9 mammography technologists. Observers interacted with the display to decide whether each image contained either no malignant lesions, or, suspicious lesions indicating malignancy. Decision time was measured as lesions were localized, classified and rated for decision confidence.

Results: Compared to experts, AFROC performance was significantly lower for residents and equivalent to technologists. Net result of analysis of overall performance was that as level of expertise decreased false positives exerted a greater impact on overall decision accuracy over the time course of image perception. This defines the decision-speed accuracy relationship that characterizes mammography expertise.

Conclusions: Differences in resident performance resulted primarily from lack of perceptual-learning experience during mammography training which limited object recognition skills causing competition between true malignant lesions, benign lesions and normal image perturbations. A proposed solution is systematic mentor-guided training that links image perception to feedback about reasons underlying decision making.

Keywords: Mammography Expertise, Perceptual and Cognitive Skills, Breast-Lesion Detection AFROC Performance, Decision Time, Feedback Training in Mammography

INTRODUCTION

One of the outstanding characteristics of an expert in radiology is the speed and accuracy of deciding whether an abnormality is present on a medical image (1,2,3). Acquiring expertise in radiology requires specialized training, experience and some degree of talent. How much and what kind of training and experience has been the subject of an organized body of research that has emerged from the field of artificial intelligence (4,5). The present paper explores the roles of training and experience on the development of expertise in mammography by comparing the performance of experienced radiologists (mammographers), radiology residents and mammography technologists. Our study focused on the performance of the radiology residents who were receiving training and mentor-guided experiences during mammography rotations that presumably provide a basis for mammography expertise.

It is difficult to find a yardstick to quantify the experience required to achieve expertise in mammography, but one could consider a reading on each case that results in a diagnostic report as a learning-experience trial. This ignores immediate feedback, which is important for perceptual learning, but is typically absent in clinical practice. In the context of medicine, training consists of mentored experience in which the resident reads medical images and then reviews them with the mentor. This training is designed to build feedback into the mentor-guided reading experience, but feedback is neither immediate nor systematic once the resident enters practice. If for the moment each read-and-reported case is considered an experience trial regardless of whether or not feedback accompanied it, it has been estimated (6) that expertise in mammography translates roughly into an average reading experience equivalent of about 10,000 cases over a period of three years. This amount of experience compares favorably with estimates of the number of games a chess player plays to reach grandmaster status (7). The average radiology residents' case-reading experience in a mammography rotation over 4 years is about 650 cases, of which only a dozen or less may be actual cancers. This means that extensive reading experience after residency will be required to reach proficiency as a mammographer. Thus, the amount of experience that a radiology resident receives in training is literally a drop in the expertise bucket.

Visual search is important for detecting lesions in mammograms, but this search skill in experts seems to be specifically tuned for detecting breast lesions embedded in mammograms, and does not transfer to similar search tasks in which hidden words and figures are embedded in pictorial scenes (8). It may not even effectively transfer to reading x-ray images outside of the breast. Efficient search skills make expert mammographers fast, accurate recognizers, classifiers and decision makers. Eye-position studies have shown that experts are faster detecting lesions in chest or breast x-ray images than less expert observers, and that visual-gaze duration (dwell) which is assumed to reflect visual information processing is related to decision outcome (6,9). In general, observers dwell longest on areas where they report abnormalities, either true positives or false positives. Areas that are considered negative receive the shortest dwell times. False negative decisions are the exception. In many instances, observers dwell almost as long on areas that contain abnormalities but are reported negative as they do on similar areas that are reported positive suggesting that the area was troublesome even though reported negative.

The fact that cumulative dwell predicts misses is important in the context of the present study, because it reflects recognition and decision making leading up to a diagnostic outcome in much the same way that decision time reflects the gathering of information leading up to a localization decision. However, visio-spatial localization of regions of interest obtained by eye-position recording cannot be derived from decision-time data. The analysis of visual dwell and its relation to information processing leading to a decision outcome suggests that chronometric analysis of the relationship between decision times and decision outcomes may compliment visual dwell data. Experimental psychology has studied reaction time, which is closely related to decision time in the present study, because it "...can help one trace the time course of information processing in the human nervous system" (10, p. 218).

If the goal of mentor-guided experience during resident training is to provide the basis for expertise in mammography, then an important question is: What kind of skills are acquired? Previous research has helped to identify three general areas in which experts skills operate (a) visual search, (b) pattern and object recognition, and (c) decision making. Since a key characteristic of mammography expertise is speed-accuracy relationship in decision outcome, the present study will focus on how decision making changes as a function of training and experience by comparing groups of observers differing on speed and accuracy dimensions. This entails measuring decision times of observers during mammographic interpretation on a digital workstation and analyzing their decisions by comparing them against a truth table. Three questions will be explored. First, how does performance change as a function of mentor-guided reading experience. Second, how does decision outcome relate to decision time for each decision event during image perception? Finally, what is the likelihood of true vs. false decision outcomes over the time course of image perception and decision making? This last question was first addressed by Christensen et al. (11). They were interested in the relationship between what they called search time and perception in the interpretation of subtle abnormalities and nonpulmonary lesions in chest radiographs. Search time was defined as how long it took to identify an abnormality, and since there was the possibility of multiple abnormalities per image, there could be multiple decisions per image. Each decision was timed and counted as a decision event. Maximum search time per image was 4 min., but most decisions took from between 1.84 and 2.68 min. on average. To compensate for the efficiency associated with faster search times, actual search time was adjusted by covarying it with the number of decision events within the maximum allotted search time per image. So experienced readers (faculty radiologists) made significantly more decisions in less time than inexperienced readers (radiology residents). By mapping the search times of decision events against a truth table they were able to plot the time course of true- and false-positive decision outcomes. The analysis of time-perception data revealed that true positives outpaced false positives throughout the time course of viewing for experienced readers, whereas false positives overtook true positives during the time course of viewing for inexperienced readers.

MATERIALS AND METHODS

The mammography test set consisted of craniocaudal (CC) and mediolateral oblique (MLO) paired views from 78 unilateral mammogram cases for a total of 156 images. The images were digitized on a Lumiscan Model 100 digitizer (Lumysis Inc., Sunnyvale, CA) using a 100 micron spot size. The mammograms were of a single breast selected by two mammographers (SO and DS) from a database of mammography cases taken from the archive of the Hospital of the University of Pennsylvania. These mammographers were later used in the study, but more than two years had elapsed prior to their testing, and each mammographer contributed only about half of the mammograms to the test set. The mammograms were assembled from cases classified by mammography assessment as normal for at least 2 years, cases classified by mammography assessment as benign with biopsy proof of benignancy and cases classified by mammography assessment as malignant with biopsy proof of malignancy. The test set contained 25 cases with 15 instances of malignant masses and 14 instances of malignant calcifications showing on both views, one instance of an architectural distortion underlying malignancy on both views of one breast and one instance of a single malignant calcification present on only one view. There were 24 cases with 12 instances of benign masses and 12 instances of benign calcifications showing on both views and 26 cases considered normal. In addition there were three practice cases: two showing lesions on both views and one lesion-free normal. Two mammographers (SO and DS) selected mammograms cases containing subtle benign and malignant lesions. Many of the normal mammograms contained ambiguous image perturbations and thus were considered "difficult normals" by the two mammographers.

The test set was displayed on a single 19-inch gray scale monitor (GMA 201, Tektronix, Beaverton, OR) interfaced to a Sun Sparc 10 computer (Sun Microsystems, Sunnyvale, CA). At the time of testing the brightness of the monitor was 127 cd/m^2. This brightness level is low for current state-of the-art mammography displays, and may have led to higher than normal error rates, at least for inexperienced viewers. Each display consisted of two views of a single breast displayed in the center of the monitor at 2048 x 2048 resolution. The gray scale was adjusted for each image by the experimenters (CFN, HLK) to a setting that covered the gray-scale range of the breast-only portion of the image. The CC view was shown on the left half of display screen and the MLO on the right half of the display screen. This is not a typical format for reading mammograms, but we were interested in determining how well observers with different levels of expertise could locate lesions in two views.

Three groups of observers representing different levels of mammography training and reading experience participated: staff mammographers with more than 5 years experience as dedicated breast imagers (n=3); second, third and fourth year radiology residents undergoing a mammography rotation (n=19); and, radiology technologists having 1-9 years experience in mammographic imaging, but no reading experience (n=9).

The procedure for testing observers was similar to the interruption technique used by Berbaum et al. (12) to obtain response times during visual search. However, the observers viewed the images on a workstation. Lesion identification and decision confidence was entered by "clicking" with a mouse-driven pointer on a menu called up at the time that a lesion was localized, and the time from the onset of the display until a decision was made, referred to as decision time, was automatically recorded. The observers were told that they were being tested on their ability to screen for malignancy in a two-view mammographic display of a single breast. If a malignancy was detected, they were to move the cursor to the lesion location and click on it. This action recorded the lesion location and called up a special menu from which they could classify the lesion as mass, calcification or architectural distortion, and rate their level of suspicion of malignancy: definitely malignant; highly suspicious for malignancy; moderately suspicious for malignancy; and, low suspicion of malignancy. If the 2-view mammogram display was determined to be free of malignancy, then the observer clicked "Return to Routine Screening" on the general menu. If a benign lesion was detected, the observer was instructed to treat it like a lesion-free image and click "Return to Routine Screening". In addition to these instructions the observers were told to localize malignant lesions on both views if possible, and to point to the center in localizing masses and center of a group of calcifications. After three practice trials with the experimenter to familiarize themselves with the workstation cursor operations, observers were left to view the 75 case test set on their own. Viewing time per case was unlimited. Decision times were recorded each time a lesion was localized by cursor control, but the observers were not told that their responses were being timed. Since multiple responses were made per two-view image pair, each localization event signaled the occurrence and time of a decision indicating the presence of a true or false malignant lesion. Figure 1 shows how these events were translated into decision-times measures. For our analysis of decision times, we used the method of survival analysis to generate the cumulative time course of decision outcomes during the time course of viewing. Survival analysis has the advantage of adjusting individual decision times for decision outcomes per case by the total decision making time required for a case. Thus, our analysis of decision times focused on the cumulative number of decision events per group over the time course of viewing. This is similar to the Christensen et al. analysis which focused on the cumulative number of decision events per group over the time course of viewing 100 chest films.

(Figure 1 here)

Analysis of Decision Time and Performance

Analysis of cursor events for localizing, classifying and rating lesions was accomplished by comparing the observers decisions against a truth table. The truth table was generated from a combination of mammographic assessment by two of the authors (SO and DS) and biopsy information on each case. Because all pairs of positive images contained at least two lesions, Alternative Free Response Operating Characteristic (AFROC) analysis was carried out treating the pair of positive images as the unit of analysis. This was consistent with the instructions for the task, and provided evidence on how well observers with different levels of mammography expertise coordinated lesion localization in a second view given lesion detection in the first view.

For the AFROC analysis, 30 pairs of malignant lesions were identified as appearing on 25 image pairs. These 60 lesions were counted in the malignant-positive category. The 24 image pairs containing benign lesions plus the 26 lesion-free images (total of 50 image pairs) made up the non-malignant category. In the AFROC analysis we counted all correctly localized lesions within plus or minus .41 cm of true location on the malignant two-view image pairs (2 standard deviations of mean accuracy of .28 cm for mammographers), and counted only the highest-rated false positive for the 50 non-malignant image pairs (equivalent to counting false-positive images or FPIs, see 13). It should be noted that this performance criterion ignores classification information which we felt unreasonably stretched the assumptions underlying the 2-Alternative Force Choice experimental framework. Basically, AFROC was designed to measure detection performance. However, because of the importance of the classification decision in mammography, we will provide a separate analysis of the classification data to show how this performance criterion is influenced by level of expertise.

RESULTS

Overall Performance

Overall detection and localization of breast lesions was assessed as a function of level of expertise. We compared the area under the AFROC, A1, for mammographers, residents and radiology technologists. The AFROC, alternative free response operating characteristic curve, plots the fraction of actual target locations reported (true positives) against the fraction of images with any false positives. In our case we plotted only the the highest-rated false positive on a normal or benign image. Figure 2 shows AFROC curves for the three groups. The average area per observer derived from analysis of variance of A1 values was for mammographers .840 (.039), for residents .653 (.058) and for technologists .592 (.062). All of these are above chance performance which for AFROC is .000. Analysis of variance of A1 values indicated, not surprisingly, that mammographers were significantly better in overall performance accuracy than either residents or technologists who did not differ from one another (p<.01, Scheffe test). By contrasting performance for these groups representing different levels of training and experience, we hoped to gain insights into the nature of mammography expertise.

(Figure 2 here)

Relation of Case Reading Experience to Development of Mammography Expertise

In order to provide a clearer picture of how the three groups differ in experience reading mammograms, we obtained data on the number of mammographic reports generated by the residents and mammographers. The 19 radiology residents who were part of the study represented mainly third-year (n=7) and fourth-year (n=8) residents plus 4 fellows with mammography reading experience varying from 10 to 2,465 cases over a 3-year interval. Over the same period the 3 mammographers read 9,459 to 12,145 cases. The relationship between A1 and log number of cases read is shown in Figure 3 for all observers. Figure 3 shows a significant linear-regression fit of the data (R^2= .667) with a positive slope suggesting that reading-skill as reflected by A1 performance increases directly with log case-reading experience (F (1,22)= 44.15, p<.0001). The regression line intercepts the y axis at A1= .293 which implies close to chance performance with zero reading experience. A log scale was used to represent the effects of case reading experience because several investigators have suggested the relationship between practice and learning is best expressed by a power function (14). The range of case-reading experience in Figure 3 was from 1 log case reading to 4.1 log case readings or about 10-12,000 cases. Two residents at the beginning of mammography training with little case reading experience performed at an A1 of about .500. The fact that their performance is above chance at the beginning of the mammography rotation can be attributed to talent, and sub-specialty training in other areas of radiology. The training levels of the observers is indicated by the numbers or letters associated with the data points. The number 1 indicates first- and second-year residents, 3 indicates third- and fourth-year residents, f indicates Fellows and m indicates mammographers. Overall performance increases in an orderly progression with training level.

(Figure 3 here)

Identification of Lesions in Two Views

Our hypothesis was that one aspect of performance that might differentiate levels of expertise was how successful observers were at identifying pairs of lesions in a two-view (CC and MLO) display. This hypothesis was based on the assumption that when mammography experts detect a lesion in one view, they look for confirmation in a different view. Mammographers talk about using projective geometry principles to predict from the detected lesion to a likely "plane of interest" in which to search for the corresponding "depth" lesion projection. If a detected lesion can be paired in a second view, this provides confirmation that it is a real target. To follow up on this, we analyzed malignant lesions (true positives) and benign lesions (false positives) that appeared on CC and MLO views per case by referring to the truth-table. The identification of paired localizations on lesion-free areas of images (false positives) was more speculative since these were imaginary. To account for paired localizations on lesion-free areas of images (false positives), we identified sequential decisions from CC to MLO view or vice versa that were classified as being malignant and of the same type (e.g. mass, calcification or architectural distortion). Consistent with the pattern of results in the AFROC analysis, the identification of paired lesions was related to level of expertise. Proportionally more paired lesions were reported, and correctly classified, for mammographers than residents or technologists. The proportion of correctly paired-lesions was .82, .56 and .50 for mammographers, residents and technologists respectively. Proportionally fewer lesions were seen and reported correctly in only one view by all groups, and the corresponding proportions were much lower: .14, .14 and .12, respectively.

Decision Time and Decision Outcome

The regression plot in Figure 3 shows the relationship between performance and case-reading experience. We hypothesized that the decision-speed accuracy relationship which is a hallmark of expertise should accompany this improvement in performance and so we looked at decision times as a function of decision outcome again taking into account that observers were interpreting a image pair containing CC and MLO views and thus possibly making two or more decisions per case. Paired decisions were broken down into those occurring to CC view on left side of display and MLO view on the right side of the display to identify the sequencing of decisions per case. For these paired decisions, decision times to the first decision were inversely related to level of expertise with mammographers significantly faster than residents (p<.01, Scheffe), and residents significantly faster than technologists (p<.0001, Scheffe). For mammographers compared to residents, 32% more of these first responses were TPs and they were reported faster than residents. Mean decision time for the first correct decision per pair was 15.66 sec.v. 21.56 sec., t (376)= 3.91, p<.001. Technologists detected fewer TPs and took even longer to decide (28.08 sec.). Decision time was also inversely related to level of expertise in a similar pattern for classification of localized lesions. Mammographers correctly classified 38 per cent more lesions and did so faster than residents (p<.05, Scheffe), and technologists (p<.001). Mean decision time for mammographers was 16.51 sec. for classifying masses and 19.77 sec. for classifying calcifications. Both of these findings support the decision-speed accuracy relationship associated with expertise.

Finally, to provide some perspective on how TP related to false negatives (FN) we looked at decision times when all lesions were completely missed on images containing malignant lesions. In this case, total image duration was assigned as the decision time. This might be considered a "clean" miss in that no lesion was reported even though it was present during the entire time that the image was examined. There were 51 percent clean misses out of 579 total false negatives, and there was little difference in mean decision times for this clean-miss FN category, ranging from 38 to 46 sec. However, the standard deviations of the mean decision times ranged from 4.6 sec. (for mammographers) to standard deviations of mean decision times between 41.6 and 52.5 sec. (for residents and technologists, respectively) indicating much indecision in failing to make a positive report after examining two views of an image containing a truly malignant lesion in these latter two groups. The range of mean decision times for clean misses was longer than any of the other decision outcome categories and seems to complement the finding of prolonged visual dwell FNs obtained from monitoring eye position. Observers spent a longer time deciding to call a positive case negative. Overall, clean-miss FNs were significantly longer than TNs (t (864)= 4.22, p<.001). Of course, we cannot confirm that the true lesions were actually scrutinized from the decision time data, but the long decision times and wide variances suggest much uncertainty surrounding decision making.

Relationship of Decision Time to Use of Confidence Ratings

The similarity of the relationship of decision outcome to decision time for mammographers and residents suggests that they may be using similar underlying detection and decision strategies.

One measure that reflects underlying decision strategy is how observers used the confidence ratings in making decisions. It is reasonable to assume that the more sure observers are that they have detected a lesion, the faster they are at making a decision. Figure 4 shows the relationships between decision time and use of confidence ratings for the three levels of expertise. The general pattern for the mammographers and residents was that decision times were inversely related to confidence rating. The longest decision times were to definitely lesion-free images (rating=1) and the shortest decision times were to definitely malignant image locations (rating=5). This pattern suggests that both groups had a similar perceptual thresholding basis for decision which is an important factor in developing a decision-making strategy. The pattern for technologists showed virtually no relationship between decision time and use of confidence ratings. Only confidence 1 ratings were prolonged, and there was no evidence of faster decision times as technologists increased their confidence rating that a malignant lesion was present on an image.

(Figure 4 here)

Time Course of Decision Outcomes

So far, two interesting generalizations come out of the decision time analysis. First, the decision-speed accuracy relationship was found to be related to level of expertise. Figure 5 summarizes the decision-accuracy relationship expressed by d' (cumulative) as a function of viewing time for mammographers, residents and technologists. Cumulative values for true positives and false positives to both normal and benign images on a per case basis (paired decisions) as a function of decision time were obtained from Survival Analysis. These values were then transformed using the formula d'= z (TP/30) - z (FP)/50) where z can be interpreted as a deviate of the unit normal curve. The formula can be thought of as correcting the true-positive fraction by the false-positive fraction. Decision accuracy consists of detecting perturbations in images, testing them for signs of malignancy, and correctly classifying them as masses, architectural distortions or calcifications. This complex decision requires discriminating malignant from benign lesions, and, malignant from normal anatomic variants in the breast image. Decision accuracy can be expressed as A1, the area under the AFROC curve, or as d', the index of detectability that is derived from the true positive fraction and the false positive fraction at a specific decision threshold as is shown in Figure 5. Looking at performance this way shows clear differences as a function of level of expertise.

(Figure 5)

Second, decision times were longer for false than true decision outcomes. We next consider whether these false decisions tended to occur early or late in the time course of image perception. We looked at both paired and single decisions. A paired decision is one in which the observer sequentially localized a suspected lesion (true or false) on both CC and MLO views. Figure 6 shows the mean number of paired true-positive decisions (TP) and paired false-positive decisions for normal regions of the images (FPN) and benign lesions (FPB) for mammographers, residents and technologists as a function of viewing time per case. Figure 7 shows the same plot for single decisions as contrasted with paired decisions. The most striking feature of Figure 6 is the high rate of FPNs relative to TPs for technologists for paired decisions, and in figure 7 the high rate of FPNs for all groups for single decisions.

(Figure 6 here)

(Figure 7 here)

These plots show for mammographers that the rate of TP decisions is faster than FPN decisions, but FPN continues to plague performance throughout the time course of viewing. The FPB decisions drop out relatively early, and thus overall performance continuously improves with decision time until about 60 sec. Perhaps our mammographers should have considered stopping at this point because FPNs increased faster than TPs. The rate of TP decisions is slower for residents due to continuous competition from FPN up to 60 sec. As with mammographers, the FPB peaks earlier. The technologists show a decrease in overall performance over time because they continue to make more FPN decisions than TP decisions.

DISCUSSION

Understanding the Nature of Expertise

The goal of the present study was to better understand the nature of expertise in mammography. Expertise in mammography as we have defined it in this paper refers to diagnostic performance skills that enable the observer to localize a breast lesion and correctly decide that it is malignant or not on the basis of two views. Admittedly, our task was somewhat artificial in the sense that we mixed lesion detection which is the focus of mammography screening with diagnostic interpretation which is the focus of diagnostic follow up. The next step is to break the task apart and do it as a two-part test which will come closer to the BI-RADS format. Moreover, even though the diagnostic skills that we are studying are an essential part of mammography diagnosis, they are quite limited as only CC and MLO views were shown with no capability for prior studies or additional views, or magnification. Ordering additional special mammographic images such as spot compression or magnification views, and performing breast ultrasound which are an important part of mammography expertise were untapped in the present study. On the basis of the information these provide, the mammographer may decide the finding is normal, benign, or probably benign but recommend short-term follow up, or biopsy.

Why are Experts Faster and More Accurate?

Our analysis has related A1 and d', measures of overall performance, to decision time in order to shed light on basic perceptual and decision-making skills. Differences in speed-accuracy between mammographers and residents seem to be related to the experience factor required to gain expertise as we have shown in Figure 3. This suggests that experts are more perceptually sensitive in recognizing lesions than those with less expertise as a result of having read more mammogram cases, seen more lesions and differentiated more lesions into malignant and benign categories. In practical terms this means that through massive amounts of experience experts became perceptually tuned to recognizing familiar common breast structures and detecting odd or novel variations in them. Three to five years of dedicated experience reading mammograms, impacts on perceptual learning by exposing mammographers to a wide set of breast-image configurations that represent most variations of normality and abnormality. We hypothesize that this concentrated case-reading experience with mammographic images impacts on perceptual learning by producing enhanced recognition skills akin to those attributed to chess grandmasters who, according to one estimate, are capable of recognizing on the order of 50,000 different chess configurations (7). It is unclear whether enhanced object-recognition skill is the result of the development of what the artificial-intelligence community refers to as chunking or template-retrieval structures that aid short-term and long-term memory (14), or as we have suggested, more critically tuned visual recognition as the result of learning and refining distinctive-feature information used to recognize deviations from prototypic normal breast structures (15, 16).

Supporting the tuning of visual recognition argument, Sowden et al. (16) have shown that massed practice detecting calcifications in positive-contrast mammograms (bright target on dark background) positively transfers to a new task in which the calcifications are displayed in negative-contrast mammograms (dark target on bright background). This suggests that perceptual learning improves perceptual sensitivity in the detection of low-contrast targets. Massed practice was defined as a detection trial followed immediately by feedback about the correctness of observer's response. This improvement in perceptual sensitivity occurred even though the amount of massed practice was limited to 720 trials followed by the transfer test. The key to improvement may be the feedback. Generalizing the Sowden et al. results, one can not help but wonder if the effects of reading experience would be facilitated by computer-assisted visual feedback about decision outcomes delivered for some subset of test cases in which truth could be verified, or at least agreement consensus reached. The purpose of systematic visual feedback is to make image perception and decision making an integral part of a perceptual-learning reading experience (6,17).

Expertise, Chest Radiology vs. Breast Radiology

In interpreting performance differences we have to be careful to separate studies of expertise in chest radiology from those in mammography because chest radiology studies have emphasized the importance of input from peripheral vision in detecting pulmonary lesions. Peripheral vision is important during search for inconspicuous pulmonary lesions because there are many anatomic landmarks in a chest radiograph (e.g. heart, ribs, lungs, diaphragm), and it has been suggested that these anatomic landmarks act as a map helping peripheral guidance of search (18). Anatomic landmarks are few in the breast (e.g. nipple and pectoralis muscle), and breast structures that might serve as landmarks (e.g. blood vessels, ducts) are interwoven into the breast image to create texture differences that are probably too subtle to be selected by peripheral vision during search. As a consequence, rather than landmarks, we believe that perturbations in parenchymal structure caused by compression of the breast during imaging and desmoplastic reaction from a growing tumor provide focal points-of-interest during visual search. The superimposition of parenchymal structures tend to make them visually conspicuous. Because superimposition of parenchymal structures has the potential to mimic breast lesions, they may be detected by peripheral vision during the initial global survey and scrutinized during subsequent focal scanning, and falsely reported as true lesions. In the detection of breast lesions, it is not only important for the observer to recognize familiar features in the image but also to recognize odd or novel features, examine these in detail (as reflected by fixations and decision time), and weight their importance in making a decision (6, 19). We assume that dwell time spent fixating the lesion, like time spent examining the image prior to making a decision, represents information processing time required to make a decision.

Decision Strategies

Our study has shown that residents develop similar decision-making strategies to those of experts. From a practical standpoint this suggests that resident training in mammography is effective in providing a general framework for learning radiology image perception skills. But residents were inferior to experts in recognizing true breast lesions. We hypothesize that this weakness is due to primarily the lack of fine-tuned visual-recognition skills. Because feedback is recognized as a critical part of the reading experience that is built into the clinical mammography rotation, it is tempting to speculate whether providing computer-assisted feedback training could facilitate visual-recognition skills and bring resident overall performance closer to that of their mentors. Despite their limited perceptual experience, many of these radiology residents will join clinical practices and read mammograms as practicing radiologists. Does this mean that diagnostic performance of practicing radiologists will suffer as a result? Probably, since the overall average performance of the residents in the present study was Az= .743 which is 12% lower than the national average of Az= .845 for 108 US radiologists assuming that the case difficulty of the two test sets was approximately the same (20).

Finally, we have showed that decision accuracy is directly related to amount of case-reading experience. At the present time many radiology departments keep track of the number of cases read by radiologists and residents, yet no recommendations have been proposed as standards.

Our data support the need for minimum requirements in terms of number of cases readings such as those proposed by the latest FDA regulations. As of 28 April 1999 this requirement is 240 case readings within last two years of residency. In addition, we believe that some less abrupt transition between residency and practice as for example double-reading experience during the first year of practice would greatly improve performance standards (23).

ACKNOWLEDGMENTS

This research was supported, in part, by Grant DAMD17-97-1-7103 between the USAMRMC and the first author.

REFERENCES

1. Lesgold A, Rubinson H, Feltovich P, Glaser R, Klopfer D, Wang Y. Expertise in a complex skill: Diagnosing x-ray pictures. In Chi MTH, Glaser R, Farr MJ (Eds.) The nature of expertise. Hillsdale, NJ: Lawrence Erlbaum, 1988:311-142.

2. Kundel HL, La Follette P. Visual search patterns and experience with radiological images. Radiology 1972;103:523-528.

3. Parasuraman R. Effects of practice on detection of abnormalities in chest x-rays. Proceedings Human Factors Society 1986;309-311.

4. Newell A, Simon HA. Human problem solving. Englewood Cliffs, NJ:Prentice-Hall, 1972.

5. Chi MTH, Glaser R, Farr MJ. The nature of expertise. Hillsdale, NJ: Lawrence Erlbaum, 1988.

6. Nodine CF, Kundel HL, Lauver SC, Toto LC. Nature of expertise in searching mammograms for breast masses. Acad Radiol 1996:3:1000-1006.

7. Chase WG, Simon HA. Perception in chess. Cognitive Psychology 1973;4:55-81.

8. Nodine CF, Krupinski EA. Perceptual skill, radiology expertise, and visual test performance with NINA and WALDO. Acad Radiol 1998; 5:603-612.

9. Kundel HL, Nodine CF, Krupinski EA. Computer-displayed eye position as a visual aid to pulmonary nodule interpretation. Invest Radiol 1990;25:890-896.

10. Posner MI. Chronometric explorations of mind. New York:Oxford, 1986.

11.Christensen EE, Murry RC, Holland K, Reynolds J, Landay MJ, Moore JG. The effect of search time on perception. Radiology 1981;138:361-365.

12. Berbaum KS, Franken EA, Dorfman DD et al. Time course of satisfaction of search. Invest Radiol 1991;26:640-648.

13. Chakraborty DP, Winter LHL. Free-response methodology:Alternative analysis and a new observer-performance experiment. Radiology 1990;174:873-881.

14. Gobet F, Simon, HA. Templates in chess memory: A mechanism for recalling several boards. Cognitive Psychology 1996;31:1-40.

15. Myles-Worsley M, Johnston WA, Simons MA. The influence of expertise on x-ray processing. J Experimental Psychology:Memory & Cognition 1988;14:553-557.

16. Sowden P, Davies I, Roling P. Perceptual learning of the detection of features in x-ray images:A functional role for improvements in adults' visual sensitivity? J Experimental Psychology:Human Perception & Performance 1999; in press.

17. Anderson JR. Cognitive psychology and its implications (4th ed). New York: WH Freeman, 1995. (304)

18. Kundel HL, Nodine CF, Thickman D, Toto L. Seaching for lung nodules: A comparison of human performance with random and systematic scanning models. Invest Radiol 1987;22:417-422.

19. Ullman S. High-level vision: Object recognition and visual cognition. Cambridge, Mass., MIT Press; 1996:161.

20. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists. Arch Med 1996;156:209-213.

21. Swetts JA, Getty DJ, Pickett RM, D'Orsi CJ, Selzer SE, McNeil BJ. Enhancing and evaluating diagnostic accuracy. Medical Decision Making 1991;11:9-18.

22. Nodine CF, Kundel HL. Using eye movements to study visual search and to improve tumor detection. Radiographics 1987;7:1241-1250.

23. Beam CA, Sullivan DC, Layde PM. Effect of human variability on independent double reading in screening mammography. Acad Radiol 1996;3:891-897.

FIGURE CAPTIONS

Figure 1. Diagram showing the relationship between image-display presentation and decision events signalled by the observer clicking the location of a breast lesion on an image with the mouse. The measurement of decision times was from the onset of the image display to the onset of a decision event. Performance was measured for the task of reading a pair of breast images consisting of craniocaudal (CC) and mediolateral oblique (MLO) views. Therefore, more than one decision event was typically timed during each image-display presentation. Offset of the display occurred when the observer clicked on NEXT IMAGE.

Figure 2. AFROC curves showing mean decision accuracy for experienced mammographers (n=3), radiology residents (n=19), and mammographic technologists (n=10). For this analysis it was assumed that there were 60 malignant lesions on 25 image pairs consisting of CC and MLO views, and that there were 50 malignant-free images. False-positives were counted only on malignant-free images. ROCFIT was performed after averaging over the confidence intervals for each group of observers.

Figure 3. A regression analysis of overall performance measured as A1 as a function of log ^{10} number of cases read over a 3-year period by 3 experienced mammographers and 19 radiology residents undergoing clinical mammography rotation. When case readings is zero, the regression line intercepts the y-axis at A1=.393 which is close to chance performance. With mentor-guided case-reading training and experience, A1 performance increases. The numbers and letters within the figure indicate the level of training of the observers: 1 = first- and second-year residents, 3 = third- and fourth-year residents, f = fellows and m= mammographers.

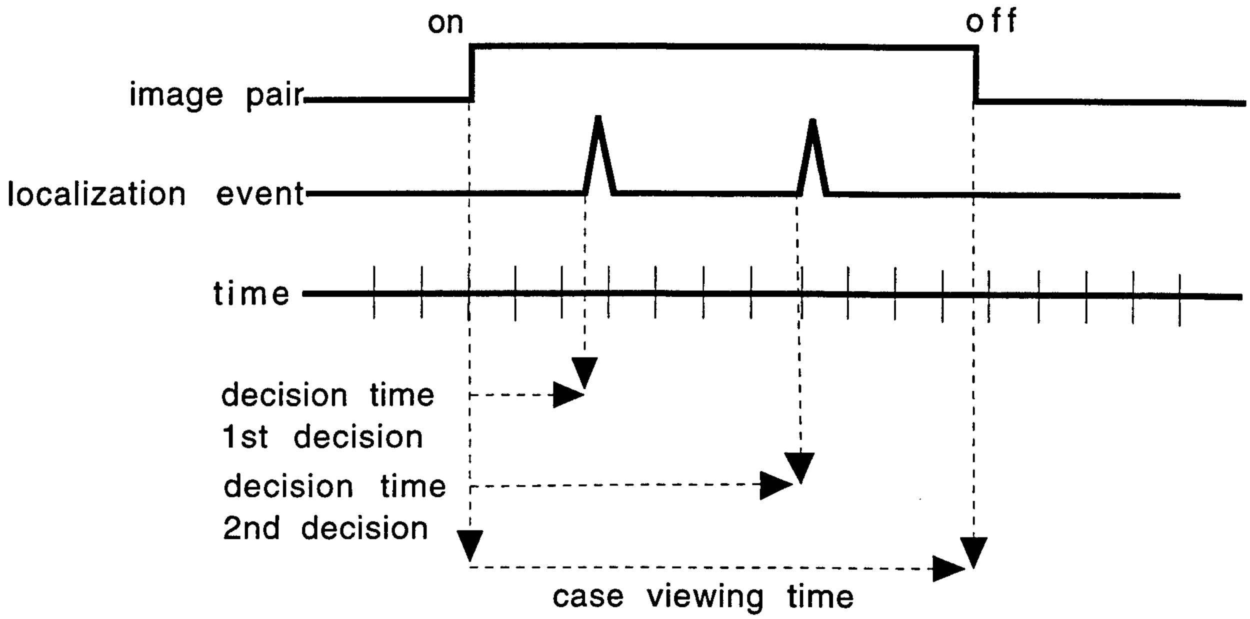
Figure 4. Decision time as a function of decision-confidence ratings for mammographers, residents and technologists. A confidence rating of 5 indicated definitely malignant, 4 indicated highly suspicious for malignancy, 3 indicated moderately suspicious for malignancy, 2 indicated low suspicion of malignancy, and 1 indicated definitely malignant-free.

Figure 5. Speed-accuracy relationship as indicated by d', the index of detectability, as a function of decision time for mammographers, residents and technologists. Overall performance as measured by d' which is the normal deviate, z, of true positives/false positives increased for mammographers and to a lesser extent for residents. Overall performance decreased below chance (d'= 0) for technologists meaning that false positives outnumbered true positives.

Figure 6. Cumulative mean number of paired decisions per case as a function of the decision time course of viewing for true-positive decision outcomes (TP), false-positive decision outcomes on non-malignant normal images (FPN), and false-positive decision outcomes on images containing benign lesions (FPB) for mammographers, residents and technologists. Paired decisions were measured. All but one mailgnant case contained lesions in both CC and MLO views. As this figure indicates, within 60 sec. mammographers had localized 23/25 or 92 percent of the paired true lesions.

Figure 7. Cumulative mean number of single decisions as a function of the decision time course of viewing for true-positive decision outcomes (TP), false-positive decision outcomes on non-malignant images (FPN), and false-positive decision outcomes on images containing benign lesions (FPB) for mammographers (M), residents (R), and technologists (T).

Figure 1


on
off
image pair
localization event
time
decision time
1st decision
decision time
2nd decision
case viewing time

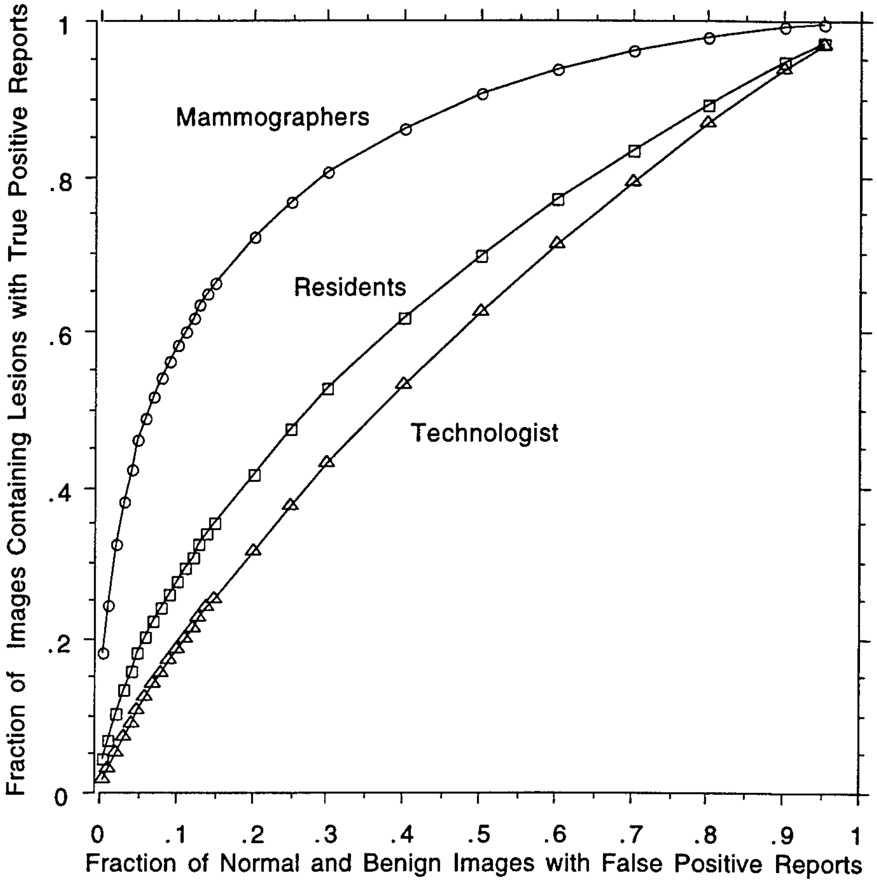

Fraction of Images Containing Lesions with True Positive Reports
0
.2
.4
.6
.8
1
Mammographers
Residents
Technologist
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
Fraction of Normal and Benign Images with False Positive Reports

Figure 2

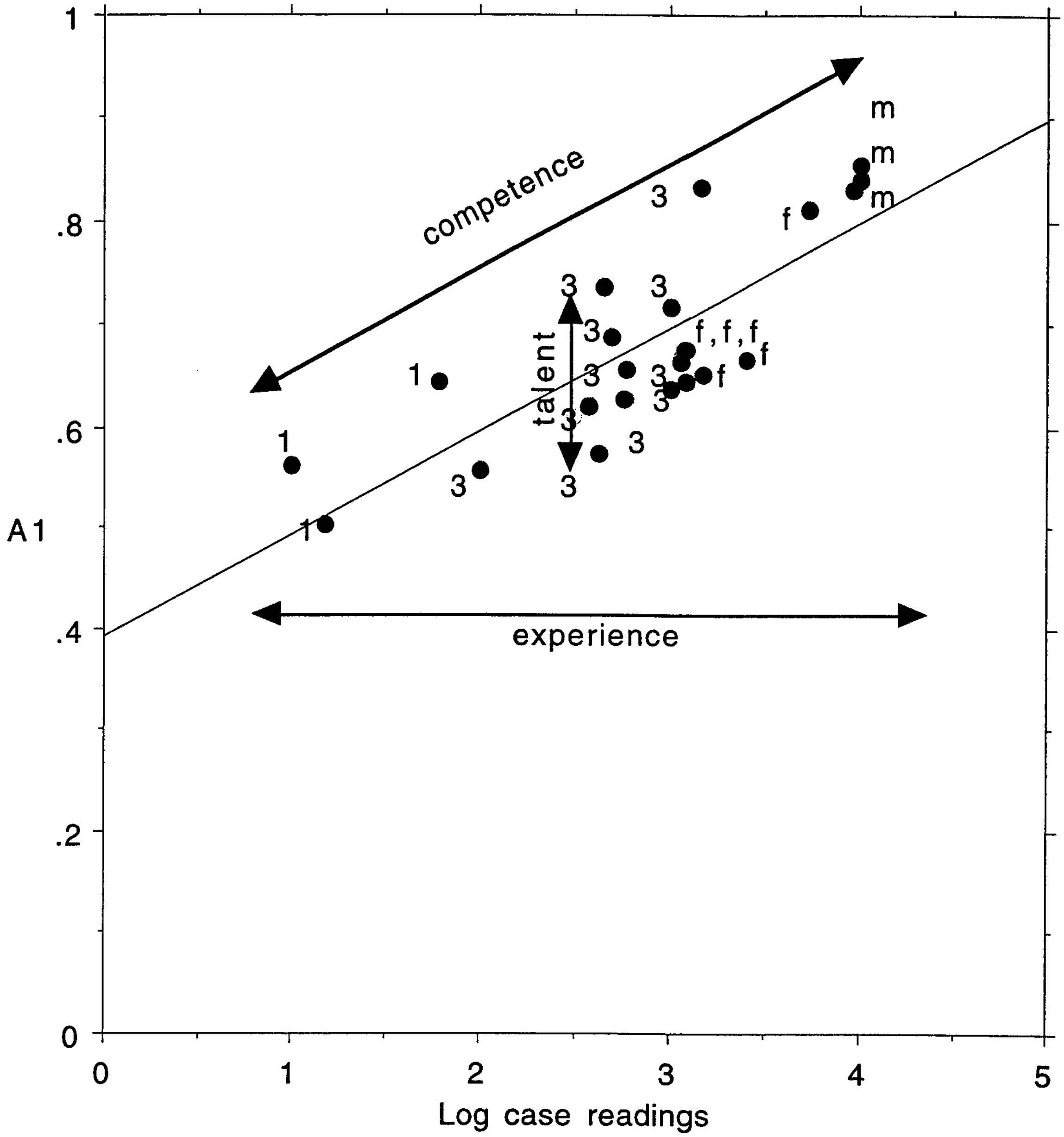

competence
talent
experience
A1
Log case readings
1
.8
.6
.4
.2
0
0
1
2
3
4
5
1
1
1
3
3
3
3
3
3
3
3
3
3
3
3
f
f,f,f
f
f
m
m
m

Figure 3

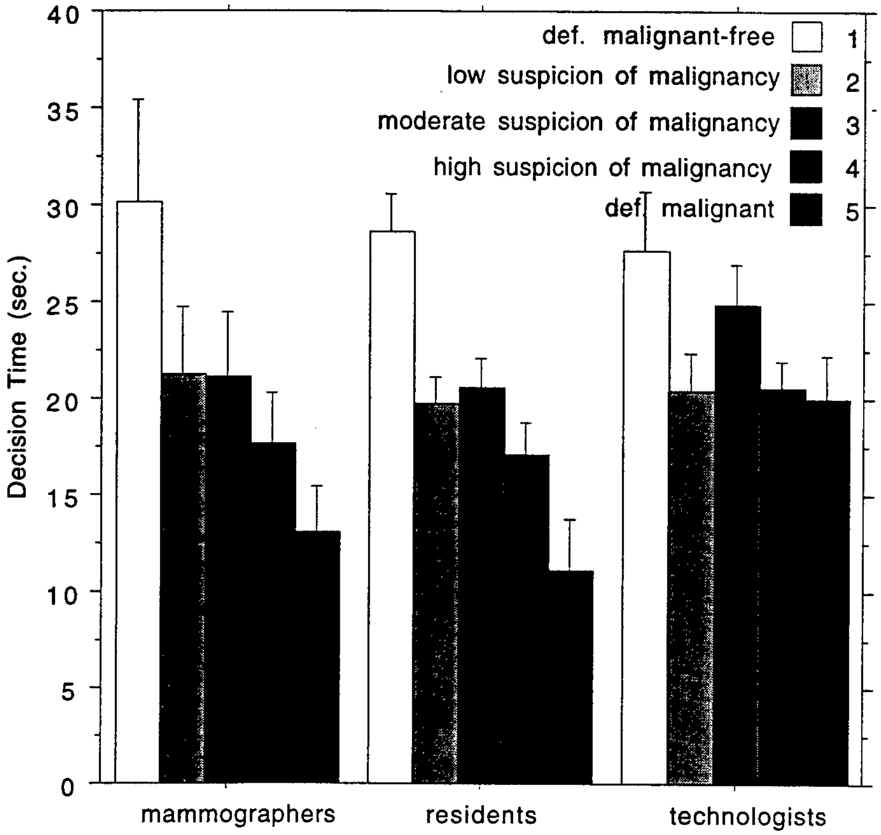

Decision Time (sec.)
40
35
30
25
20
15
10
5
0
def. malignant-free 1
low suspicion of malignancy 2
moderate suspicion of malignancy 3
high suspicion of malignancy 4
def. malignant 5
mammographers
residents
technologists

Figure 4

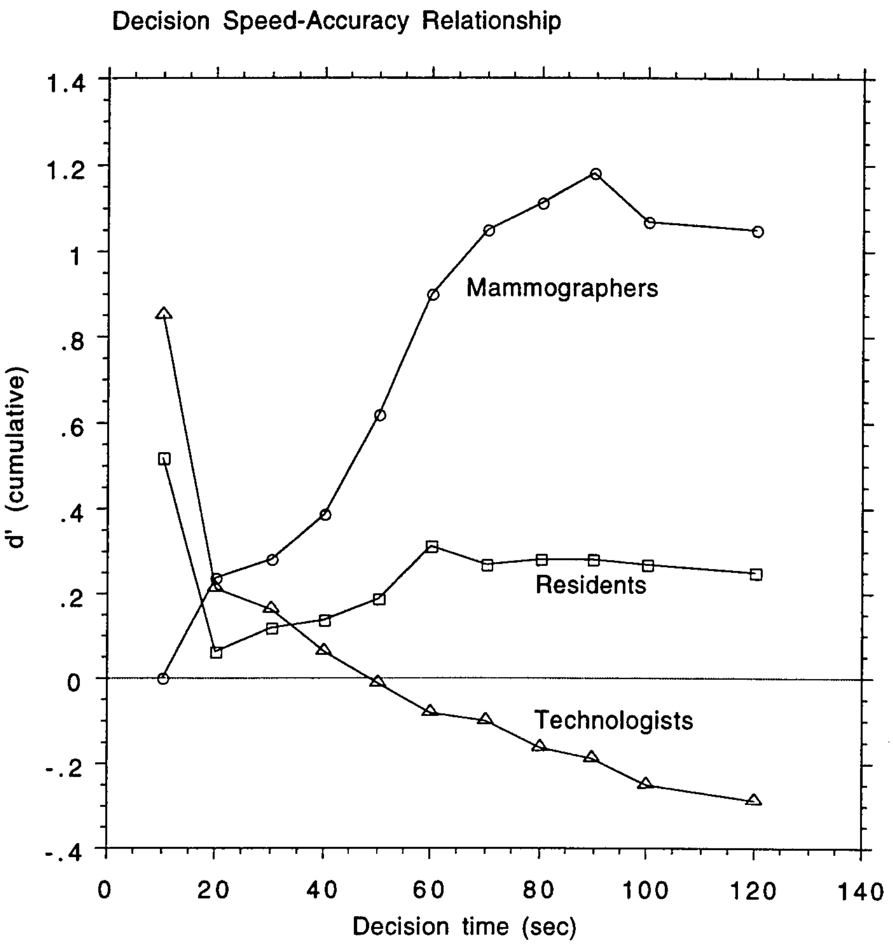
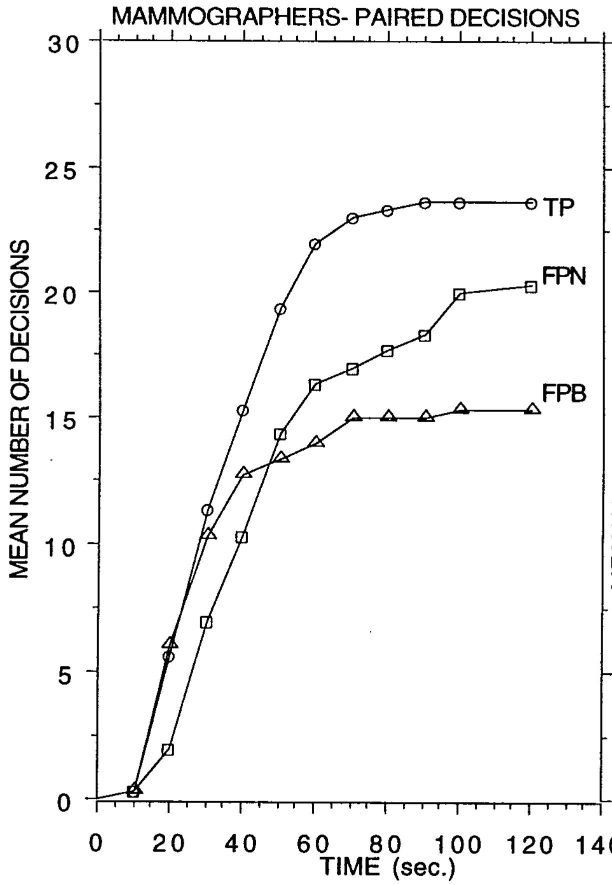
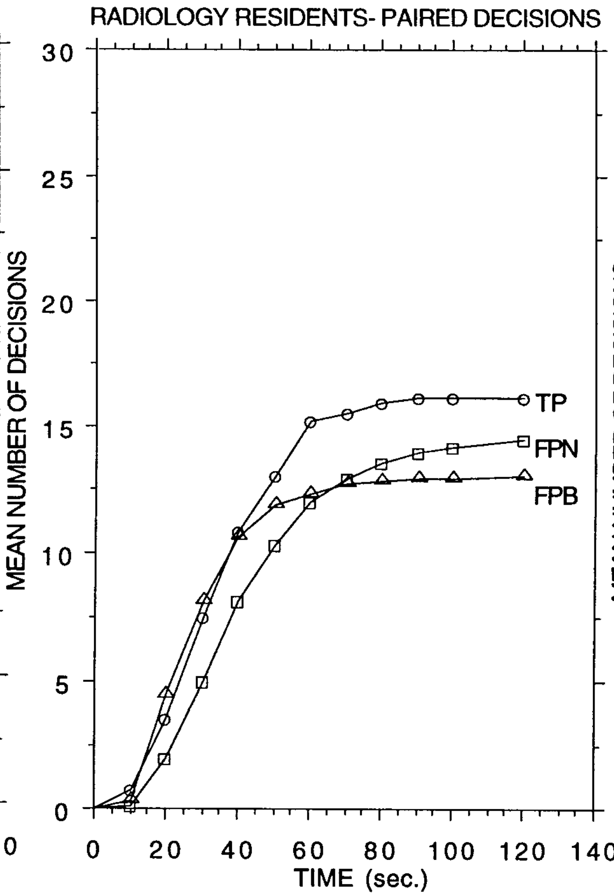

Decision Speed-Accuracy Relationship
d' (cumulative)
1.4
1.2
1
.8
.6
.4
.2
0
-.2
-.4
0
20
40
60
80
100
120
140
Decision time (sec)
Mammographers
Residents
Technologists

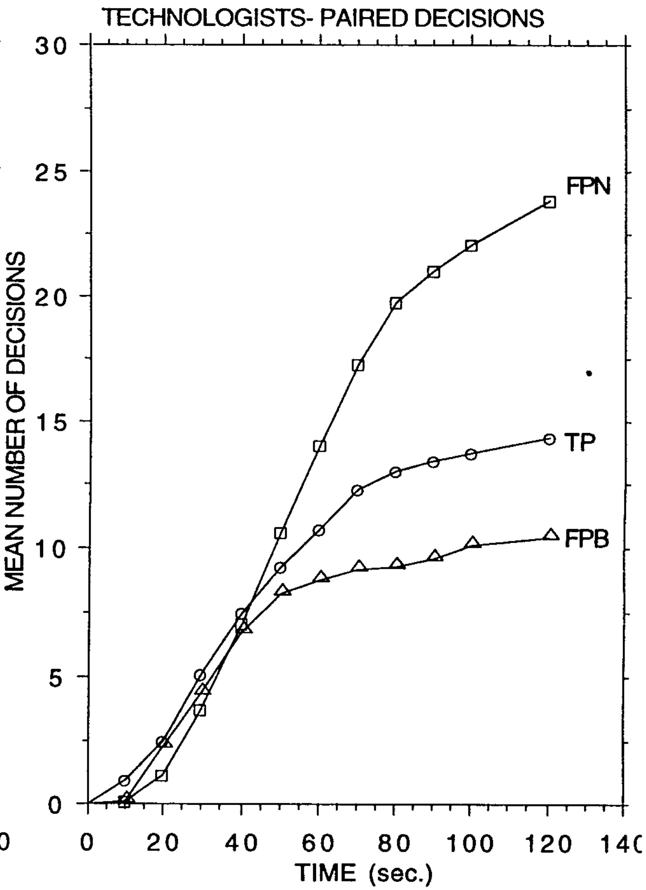
Figure 5



MAMMOGRAPHERS- PAIRED DECISIONS



RADIOLOGY RESIDENTS- PAIRED DECISIONS



TECHNOLOGISTS- PAIRED DECISIONS

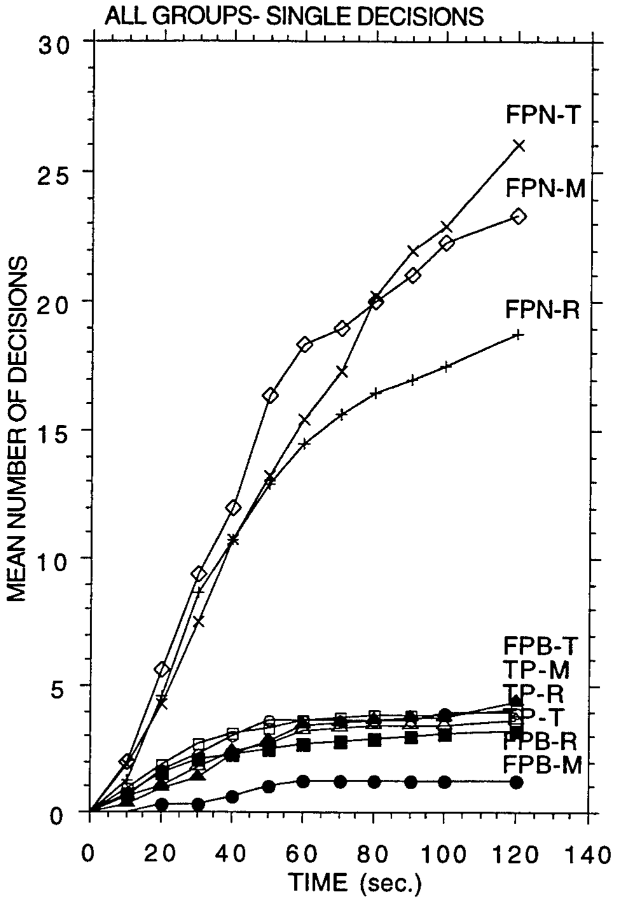

ALL GROUPS- SINGLE DECISIONS
MEAN NUMBER OF DECISIONS
30
25
20
15
10
5
0
0
20
40
60
80
100
120
140
TIME (sec.)
FPN-T
FPN-M
FPN-R
FPB-T
TP-M
TP-R
TP-T
FPB-R
FPB-M

Figure 7

Appendix 2

PROCEEDINGS OF SPIE REPRINT

SPIE—The International Society for Optical Engineering

Reprinted from

Medical Imaging 1999

Image Perception and Performance

24–25 February 1999
San Diego, California

SPIE
P
PROCEEDINGS
SERIES

Volume 3663

A Chronometric Analysis of Mammography Expertise

Claudia Mello-Thoms[1], Calvin F. Nodine and Harold L. Kundel
University of Pennsylvania School of Medicine, Philadelphia PA 19104

ABSTRACT

This paper studies the effects of training and experience on decision time and performance in mammography. We compared the performance of three groups of observers representing different levels of expertise: dedicated breast imagers (mammographers), radiology residents undergoing a mammography rotation, and mammography technologists, when reading a test set that contains benign and malignant lesions, as well as lesion free images. We show that the number of cases read significantly impacts performance, as measured by the area under the AFROC curve. We also show that different levels of expertise have different decision structures during the time course of image viewing. In fact we show that the mammographers should stop reading an image after 60-80 seconds, because at this point they have found all of the true targets present, and they are much more likely to make a mistake. On the other hand residents and technologists mistakes plague their performance throughout the time course of image viewing.

Keywords: expertise, mammography, AFROC analysis, time course of image viewing.

1. INTRODUCTION

Breast cancer is one of the leading causes of death among American women. It is estimated that one woman is diagnosed with this disease every three minutes, and from these, 46,000 will die each year (3). Many techniques are available for diagnosing breast cancer; the most widely used is Mammography, due to its cost/effectiveness ratio, which allows for the screening of large parcels of the population. It has been shown (4) that mammography screening leads to a reduction of breast cancer mortality of 29-45% in women in their forties, and 34% for older women.

In this paper the roles of experience and training in mammography expertise are studied. We compared the performance of experienced radiologists dedicated to breast imaging (mammographers), radiology residents undergoing a mammography rotation and mammography technologists, when reading a test set composed of 78 two-view mammograms containing benign and malignant lesions, as well as lesion free cases. These three groups differ in their levels of formal learning (training) and total number of cases read (experience). As a consequence, the speed and accuracy relationship, which is a hallmark of expertise, is clearly observed in the decision structures of these three groups.

Although it is almost impossible to find one measure that defines an expert in mammography, one can consider that each case read, with or without feedback, corresponds to a learning trial. In this sense it has been estimated (1) that expertise in mammography translates roughly to an average of 12,000 cases a year over a period of 3 years. If one considers that the average radiology resident sees, over a period of 4 years, around 900 cases, of which perhaps a dozen are actual cancers, then it becomes clear that many more years of dedicated work will be necessary to elevate that radiologist to the level of his or her expert peers. This paper will explore the relationship between the number of cases read and performance, as measured by the area under the AFROC curve.

Also, the features that signal breast cancer may be very small and difficult to find. This is translated in a False Negative rate of 10 to 30%, of which 2/3 are seen in retrospect (2). Because these False Negatives (FNs) may have potentially deadly consequences, experts learn to over-read the cases, which generates high rates of False Positives (FPs). When these FPs occur in the time course of decision making will also be examined in this paper.

2. MATERIALS AND METHODS

The test set used consisted of 78 image pairs representing the cranial-caudal (CC) and the medial-lateral oblique (MLO) views of the breast. It was digitized using a Lumiscan Model 100 digitizer (Lumisys Inc., Sunnyvale, CA) using a 100 micron spot size. This test set was assembled from cases considered normal for two years by mammographic assessment and cases with benign and malignant lesions that were biopsied and thus confirmed as being either benign or malignant.

[1] Correspondence: Email: cthoms@mipgsun.mipg.upenn.edu

Part of the SPIE Conference on Image Perception and Performance
San Diego, California • February 1999
SPIE Vol. 3663 • 0277-786X/99/$10.00

The test set was displayed on a single 19-inch 2048 x 2048 gray scale monitor (GMA 201, Tektronix, Beaverton, OR) interfaced to a Sun Sparc 10 workstation (Sun Microsystems, Sunnyvale, CA). Each display consisted of two views of the same breast: on the left hand side was the CC view, and on the right hand side was the MLO view.

The observers were instructed to indicate malignant lesions only. They were free to search the image for as long as they wished. Upon encountering a malignant lesion, they were supposed to move a mouse controlled cursor to the center of that malignant lesion, and click. This action would cause a window to open, in which they had to indicate the type of the lesion (mass, calcification, architectural distortion), and their level of confidence that that lesion was indeed malignant (definitely malignant, highly suspicious for malignancy, moderately suspicious for malignancy and low suspicious of malignancy). Note that the observers were instructed to indicate the same lesion in the other view of the breast, if they could see it. If only benign lesions were found, the observers were instructed to go to the next image, by clicking a button entitled "Return to Screening". Similarly, if no lesions were found the observers were instructed to go to the next image. Also, in order to get information about the different experience levels, we obtained data on the number of mammographic reports generated by the residents and mammographers.

3. RESULTS

Detection and Localization of Malignant Lesions: We assessed the observers abilities to detect and localize malignant breast lesions as a function of the observers' expertise. The area under the AFROC, A1, was used to compare the three groups. The average area per observer derived from analysis of variance of A1 values was .840(.039) for mammographers, .653(.058) for residents and .592(.062) for technologists. Analysis of variance indicated that the mammographers were significantly better (p<.01, Scheffe test) than either residents or technologists. Furthermore, these last two groups did not differ significantly from one another.

Performance vs. Experience: The 19 radiology residents who were part of our study were primarily third- and forth-year residents, and three of them were fellows at the time of these tests. They had a mammography reading experience that varied from 10 to 2465 cases over a 3-year interval. Over the same period the mammographers read between 9459 and 12145 cases. The relationship between A1 and the log (base 10) number of cases (we used the log because of the power law of learning (5)) shows a significant linear regression fit (R2=.667), having a positive slope, which indicates that case reading experience indeed influences A1 performance over a wide range of experience (F(1,22)=44.15,p<.0001). This is shown in Figure 1.

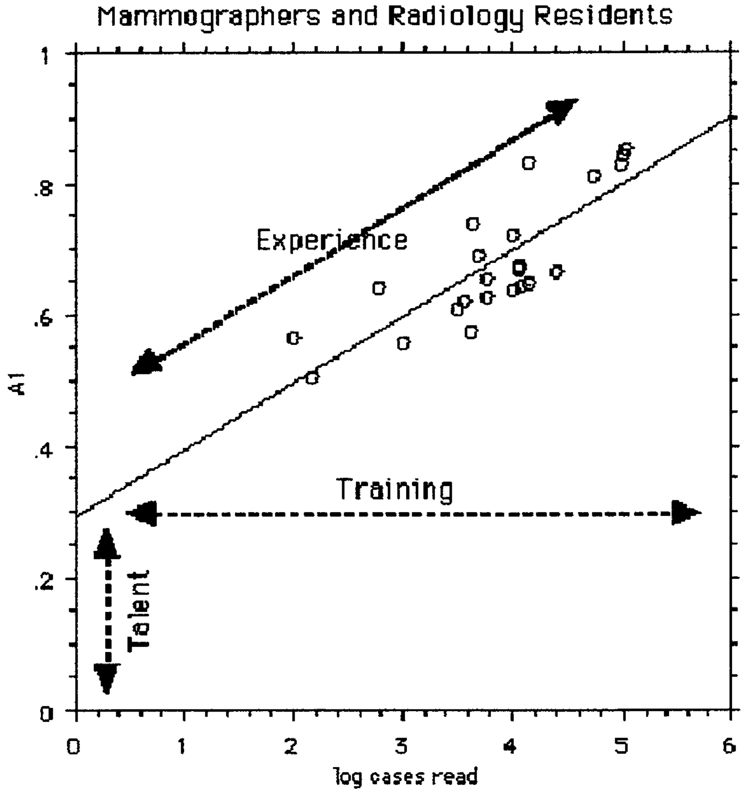


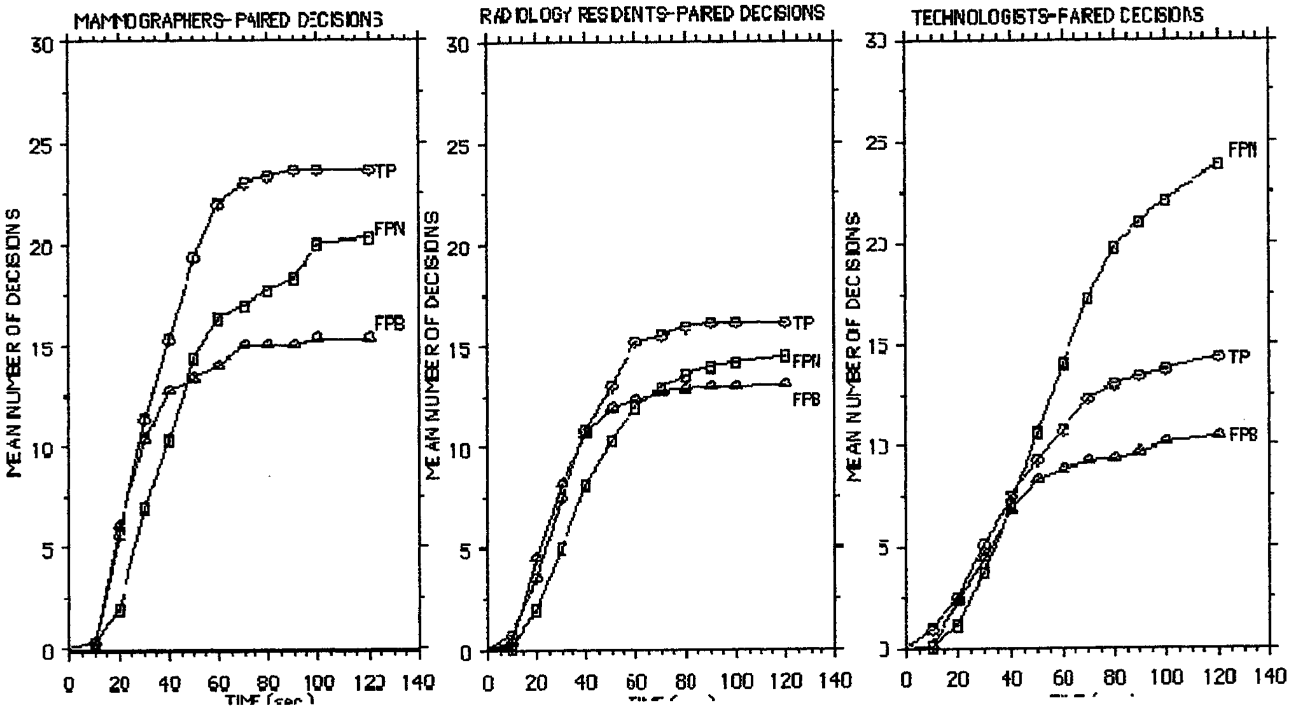
Figure 1. The roles of experience and training in observer performance, as measured by the area under AFROC, A1.

Note that the regression line intercepts the y axis at A1=.293. This implies that, with zero reading case experience, the observers performed close to the chance line, which is A1=.00 for AFROC. The fact that the performance is still above the chance line in the beginning of a mammography rotation can be explained by taking into account that these residents were already exposed to general radiology residence training.

Decision Times vs. Decision Outcomes: We also looked at the decision times as a function of decision outcome. Note that because the observers were instructed to indicate the lesion pair (that is, the same lesion on both views of the breast), when possible, they made on average two or more decisions per case (a case being a set with two images, one CC and one MLO view of the same breast, displayed simultaneously). For these paired decisions, decision times to the first decision were inversely proportional to the level of expertise. The mammographers were significantly faster than the residents (p<.01, Scheffe test), and residents significantly faster than technologists (p<.0001, Scheffe test). When comparing the mammographers with the residents we found that 32% more of these first decisions were True Positives, and they were reported faster than residents'. Mean decision time for the correct decision per pair was 15.66 sec v. 21.56 sec, t(376)=3.91, p<.001. On the other hand the technologists detected even fewer True Positives, and did so at a much slower pace (28.08 sec).

Time Course of Image Viewing: Finally we looked at the question of when the True (TPs) and False Positive (FPs) decisions occurred in the time course of image viewing. We have divided the FPs into two types, namely, the ones that correspond to a benign lesion (that is, they are due to the actual presence of a lesion in the image, except that the lesion was benign), which we called FPBs, and the ones that were due not to the presence of lesions themselves, but rather to a misinterpretation of the image features, which we called FPNs. This is shown in Figure 2.

Figure 2. Time Course of Image Viewing. The first curve shows the mean number of decisions for the mammographers as a function of time. Note that the FP curves do not overtake the TP curve for the initial 140 seconds of image viewing. The second curve shows how the residents perform under the same conditions. Note that there is much more competition among the FPs and the TPs. The final curve shows the results for the technologists. In this case the FPs overtake the TPs very early on, and this leads to fewer true lesions found and many more mistakes.



For the mammographers, although there is competition between the FPNs and the TPs during the time course of image viewing, the former never overcomes the latter, which provides further evidence that mammographers are in fact well trained target detectors. Furthermore, the FPBs are the first ones to die off, probably because, being the most common type of lesion seen on their daily practices, they are easier to discriminate from malignant lesions. For the residents the behavior is different. Although the FPNs never overtake the TPs, the competition is much more fierce. Moreover, there is also competition from the FPBs, which reflects the fact that the residents have been exposed to a smaller number of cases, and so to them it is harder to differentiate between benign and malignant lesions, as well as artifacts (that is, object-like structures that they may see in the image but that do not correspond to an actual lesion).

The technologists behavior is quite different than the other two groups. Due to a combination of lack of formal training in reading mammograms and lack of experience in doing so, they use what we chose to call 'the shot-gun strategy'. This means that anything that looks even slightly suspicious gets called as being malignant. This low threshold forces them to make many False Positive calls, and these calls overtake, very early on, the decision course of the TPs. As a consequence the technologists not only make many more calls per image but also commit many more errors, and have the lowest rate of true malignant lesions found.

4. DISCUSSION

The analysis of our data from three perspectives, namely, performance, experience and decision time, has shown that the mammographers performed significantly better than either residents or technologists when reading a test set composed of 78 two-view mammograms. We hypothesized that this better performance is directly related to the number of cases seen, because by seeing more cases they become more attuned to the features that signal a malignant lesion. They also become better at differentiating benign from malignant lesions, as well as they are less responsive to features in the image that potentially mimic a lesion, such as superposition of structures, etc. We believed that the number of cases seen has a reflection in the perceptual process, making the mammographers better target detectors and better target classifiers, in the sense that the memory schema for separating malignant from benign lesions is larger than that of either of the other two groups.

We showed the importance of experience by the significant correlation between the logarithm of the number of cases read and the performance, as measured by the area under the AFROC curve.

Furthermore, we have showed that the mammographers are significantly faster decision makers than either residents or technologists, and they are better at classifying lesions as being benign or malignant. They make many more first calls that are True Positives. Note that this reflects a speed-accuracy relationship, which is a hallmark of expertise.

In order to understand when the True and False Positive calls are made during the time course of image viewing we have studied the decision behavior of the three groups. The mammographers, being better target detectors, find most of the TPs within the first minute of image viewing. Moreover, because of their familiarity with the benign lesions, they are able to find most of these within 30 seconds of viewing the image. The interesting information in this curve is that after 60 seconds their FPNs start to rise, and eventually they overtake the TPs. This seems to indicate that the mammographers would make less errors if they stopped reading the image after 60-80 seconds, while their TPs are still dominating the FPB and FPN curves.

In the case of the residents this behavior is somewhat different, although the same pattern set by their mentors is present in that the TP curve still dominates the FP curves. In this case, however, competition is more fierce, as indicated by the longer decision times, and thus attention is very divided. The differentiation between the benign lesions (FPBs) and ones that are due to misinterpretation of image features (FPNs) becomes less pronounced, possibly as a result of having seen fewer benign lesions. Furthermore, the TPs have to constantly compete with these erroneous calls, which depresses the TP decisions.

For the technologists the situation is even worse. Their perceptual decision making process appears to have a very low threshold, causing them to call anything that seems slightly suspicious in the image as being malignant. This behavior obtains that very early on the FPs overcome the TPs, which leads the technologists to make many more errors and to find many fewer true lesions.

5. CONCLUSIONS

We have studied the roles of experience and training in mammography expertise. We have shown that not only is there a significant difference in performance as a function of level of expertise but also a difference in decision making strategies that impacts on decision time.

Our results showed that the experts are faster and better at localizing malignant lesions in a mammogram test set. This is a reflection of the speed-accuracy relationship. Furthermore, our results have shown when the experts should stop reading the films, because at a certain point in the time course of decision making False Positives overtake the True Positives. Note that this results provide support for how the experts are better, but the problem of why they are better is still a very open one.

ACKNOWLEDGEMENTS

This work has been partially supported by Grant DAMD17-97-1-7103 between the USAMRMC and CF Nodine.

REFERENCES

(1) CF Nodine, HL Kundel, SC Lauver, LC Toto. "Nature of Expertise in Searching Mammograms for Breast Masses". *Academic Radiology* 3, pp. 1000-1006, 1996.

(2) M Giger, H MacMahon. "A Perspective on False Positive Screening Mammograms". *The American College of Radiology Bulletin* 34:3, pp. 565-596, 1996.

(3) CM Kocur, SK Rogers, LR Myers, T Burns, M Kabrinsky, JW Hoffmeister, KW Bauer, JM Steppe. "Using Neural Networks to Select Wavelet Features for Breast Cancer Diagnosis". *IEEE Engineering in Medicine and Biology Magazine,* pp. 95-102, 1996.

(4) SA Feig. "A Perspective on False Positive Screening Mammograms". *The American College of Radiology Bulletin* 54:6, pp. 8-9, 1998.

(5) JR Anderson. *Cognitive Psychology and its Implications.* WH Freeman, 1995.

A PERCEPTUALLY TEMPERED DISPLAY FOR DIGITAL MAMMOGRAMS

Pendergrass Laboratory
Department of Radiology
University of Pennsylvania Health System
36th Street and Hamilton Walk
Philadelphia, PA 19104

Harold L. Kundel,M.D., Susan Weinstein,M.D., Emily Conant,M.D.,
Lawrence Toto,B.S., Calvin Nodine,Ph.D.

Supported by Grants # DAMD17-96-1-6153 and DAMD17-97-1-7130 from the U.S. Army Medical Research and Materiel Command.
Recipient of a Certificate of Merit for an InfoRAD exhibit at the 1998 RSNA scientific assembly.

Corresponding Author: Harold L. Kundel, M.D.
Pendergrass Laboratory
Department of Radiology
University of Pennsylvania
308 Stemmler Hall
3600 Hamilton Walk
Philadelphia, PA 19104

Phone 215 662 6224
Fax 215 349 5115
e-mail kundel@rad.upenn.edu

Abstract

The cathode ray tube (CRT) of a workstation for use with digital mammograms was calibrated using a photometer to produce an input-output (I/O) characteristic curve similar to the perceptually linear curve defined by a current display standard. Then, a test pattern consisting of bars of increasing intensity containing disks of decreasing contrast was used by an observer to estimate the minimal detectable contrast (MDC) at different levels of display luminance. The MDC was modeled by a parabola. The shape of the parabola was determined by the observer's perceptual responses and the range by the maximum and minimum pixel values of the breast parenchyma. As each mammogram was displayed the contour of the breast was automatically found and pixels within the breast image were sampled to determine the pixel values that were used to compute the minimum and maximum pixel values. The parabola was integrated to determined the look-up table for the initial MDC tempered display of the mammogram. Preliminary observer performance tests showed no significant differences in the speed and accuracy of 3 radiologists reading a set of mammograms on the MDC tempered display when compared with the perceptually linear display.

This article describes preliminary observer tests of a method for the initial display of a digital mammogram that compensates for the display brightness , the ambient light and the useful range of pixel intensities in the image.

INTRODUCTION

Given the present state of the art, a static cathode ray tube (CRT) display can simulate but not duplicate the image quality of a film mammogram displayed on a lightbox. The film is displayed at higher luminance, has a greater spatial resolution and has a wider grayscale range (1). However, the film captures and displays the image using a fixed set of predetermined display parameters. An adjustable CRT display can be used to explore the full range of contrast and resolution available in the digital image by using image processing such as window-level and zoom-rove. Differences in the CRT input-output transfer characteristic and in the image processing that is applied to the digital image can result in wide variation in the base-line appearance of images. In order to have identical images look alike when displayed on different CRTs, a display standard called "perceptual linearization" has been proposed (2, 3). When this standard is used, equal changes in the pixel gray scale value produce equal changes in the just noticeable difference (JND) of luminance in the image.

A display standard provides an equivalent starting place for each image but may not provide the best gray-scale transformation for a particular image in a particular reading environment. Human contrast sensitivity depends upon the average luminance of the light reaching the eye (4). Most of the light that affects contrast sensitivity comes from the displayed image, but some comes from room illumination including that which is reflected from the CRT surface. In order to maximize the contrast sensitivity of the eye, large variations in the brightness of the image can be modulated by modifying the distribution of the gray-levels over the image (5, 6). Liu and Nodine (7) using a model first proposed by Mokrane (8) have developed an algorithm that equalizes perceived contrast over the image assuming some starting level of adapting luminance. Contrast is modified in the image on the basis of the theoretical threshold-contrast curves of Heinemann (4). The workstation described here extends the work of Liu and Nodine (7) to include adjusting the gray-scale transform for ambient illumination and adjusting the mammogram image to fit the entire gray-scale range of the CRT.

THE BASIC DISPLAY STATION

The display station shown in Figure 1 uses a Gateway GP6-266 (Pentium processor) computer (Gateway 2000 Inc., Sioux City, SD) that is interfaced to an Orwin D2300L grayscale monitor (Clinton Electronics, Loves Park, IL) using a Dome Md5/PCI interface board (Dome Imaging Assoc. Waltham, MA). The computer software is written in IDL, a high level graphics language (Research Systems, Inc., Boulder, CO).

Before using the display station the video monitor was photometrically calibrated. A Tektronix Model J17 photometer (Tektronix, Inc., Beaverton, OR) interfaced to the computer was used to measure the intensity of a 10 x 10 cm. square of uniform luminance located in the center of the display surface. The intensity of the display surrounding the square was set at a luminance of 55 cd/m^2 produced by a pixel driving intensity value of 128. The luminance was measured over 17 equally spaced pixel driving intensity values from 0 (black) to 255 (white) corresponding to 1.7 to 343 cd/m^2. The photometric data were digitized, log transformed, fitted with a 4th order polynomial using a least squares procedure and displayed on the CRT along with an ideal curve. The brightness and contrast controls were adjusted until the calibrated curve visually matched the ideal curve.

Once the CRT is calibrated it only needs occasional adjustment. The shape of the I/O transfer characteristic adjusted according to the perceptually linear display standard is shown in the top half of Figure 2.

DEVELOPING THE PERCEPTUALLY TEMPERED DISPLAY

Estimating the Minimal Detectable Contrast (MDC)

The MDC test pattern, shown in Figure 3 consists of 8 horizontal bands of increasing intensity. Each band contains 8 circular disks of decreasing contrast. It was displayed for each observer prior to a viewing session. The observer's task was to choose the "least visible" disk in each band. The observer's responses are affected by the display contrast and the ambient room lighting. The contrast of each indicated disk was used to fit a 2nd degree equation, where the independent variable is the driving level of the intensity of the band and the dependent variable is the contrast of the target chosen by the observer measured in pixel driving level units. These data are used to approximate the dependence of the observer's contrast-sensitivity on adapting luminance.

Approximation of the Contrast Sensitivity Curve by a Parabola.

Heinemann (4) measured human contrast-sensitivity at different levels of adapting luminance. Examples of the relationship at two adapting luminance levels are shown in Figure 4. When the adapting luminance increases, the curves shift to the right and roughly maintain the same shape. Applying a bright spotlight to the image shifts the observer's curve to the right and increases the contrast sensitivity. Many attempts have been made to fit the curves from Heinemann's experimental data with simple equations (5). The algorithm of Liu and Nodine (7) required advanced information about adaptation level and was computationally intensive. We simplified the Liu-Nodine algorithm by assuming that a parabola could be used to approximate contrast-sensitivity at different levels of adapting luminance (see dashed lines Figure 4). The fit is reasonable at high adapting luminance levels, corresponding to dense breasts and at scene luminance levels below the adapting luminance for both bright and dark images. The fit for regions brighter than the adapting luminance is not very good when the adapting luminance levels are low. We accepted this lack of a perfect fit in order to increase the contrast in the parts of the image that appeared dark and rapidly compute the look-up table for the correction. The shape of the parabola for each observer was determined from the MDC data and the range of the parabola was computed individually for each image. For example, a dark image from a fatty breast would have a different range than a bright image from a dense breast. A different look-up table is required for each image.

The best-fit parabola is integrated and normalized to the display intensity range of the mammogram to yield a continuous, non-linear lookup table that boosts contrast in the intensity bands that require higher contrast for detection of low contrast targets. Due to dynamic range limitations of the monitor, contrast enhancements in some segments of the lookup table require contrast reductions in other segments, thus producing a contrast-tempered lookup table. The MDC lookup table is designed to equalize the detectability of equal contrast (pixel driving level) targets, regardless of the regional mean pixel intensity surrounding the targets. The advantage of redistributing the contrast in this "tempered " fashion is to provide an initial view that allows visual access to the dark regions (skinline) as well as the light regions (muscle, dense tissue).

Matching the Look-up Table to the Pixel Intensity of the Mammogram.

As each case is displayed, the maximum and minimum pixel intensity in the breast parenchyma is determined by sampling over a region that includes breast tissue out to just beyond the skinline, thus excluding the extremes of pixel driving levels due to lead markers, labels and cassette edge artifacts. This is done using a boundary detection procedure, where after applying a median filter, an intensity threshold value 5% above the background (dark level) is selected. Using this threshold, the image is transformed to a binary image and a contour is determined on the resultant image. Image intensities are then sampled on the original breast image along 30 equally-spaced lines as shown in Figure 5. The maximum and minimum pixel driving levels are then applied to the MDC corrected lookup table so that the output intensity just matches the input intensity as shown in the bottom half of Figure 2. All of the calculations and look-up table manipulations are done using a 12 bit pixel intensity scale. The scale is transformed into an 8 bit scale for display.

Displaying the Images

The CRT is photometrically calibrated as part of the regular quality assurance program. The MDC calibration is performed before each reading session with the ambient illumination set at 1.6 Lux at the location of the observer's eyes. The calibration takes approximately 15 to 20 seconds to complete. The correction of each image is done off-line prior to the test. Observers are able to use a single slider to adjust the linearity of the MDC lookup table. The slider can smoothly adjust the gamma from an linear lookup table up to a maximum MDC setting. Figure 6 shows a breast image displayed using the standard perceptually linearized display and the MDC tempered display. Notice the difference in the visibility of the skin line (shown by the solid arrow).

EVALUATING THE DISPLAY STATION

Our development cycle includes periodic benchmark testing using a sample of cases from a database of normal and abnormal mammograms where all of the malignancies and many of the benign lesions are histologically proved. Readers are shown a cranio-caudal (CC) and a medio-lateral oblique (MLO) view and asked to move a pointer on the display to any potential malignant lesion and click on the mouse. Response time from the start of viewing each case and the location of the pointer is recorded by the software. After the click, a pull-down menu appears and the reader must select one or more of mass, calcification or architectural distortion and indicate a confidence in malignancy. These data are used to compute a receiver operating characteristic (ROC) curve and determine the area under the curve. Three readers, two mammographers and a general radiologist were given the test using 75 mammograms: 25 with malignancy, 25 with benign lesions and 25 normals. Table 1 is a comparison of the area under the ROC curve. Although each reader did better with the MDC corrected tempered display, the difference is not significant when tested with a paired t-test. The time to first point out a lesion was very variable but on average was not different for the two display modes.

SUMMARY

The speed and accuracy of the tempered display function is equal to the standard linear display function when used on a moderately bright monitor (300 cd/m^2). The initial view of the image provides visual access to lighter and darker regions of display with some sacrifice to middle intensity regions. The display can be linearized by moving a single slider. This is an attempt to simplify the user interface. Development of the display station is continuing with the addition of the use of verbal commands to modify display parameters and an eye position contingent roving window.

REFERENCES

1. Blume H, Roehrig H, Browne M, Ji TL. Comparison of the physical performance of high resolution CRT displays and films recorded by laser image printers and displayed on light-boxes and the need for a display standard. Proc. SPIE: Medical Imaging IV: Image Capture and Display 1990;1232:97-114.

2. Johnston RE, Zimmerman JB, Rodgers DC, Pizer SM. Perceptual standardization. SPIE: Picture Archiving and Communication Systems (PACS III) for Medical Archiving 1985;536:44-49.

3. Blume H, Hemminger BM. Image presentation in digital radiology: Perspectives on the emerging DICOM display function standard and its application. Radiographics 1997;17:769-777.

4. Heinemann E. The relation of apparent brightness to the threshold for differences in luminance. Journal of Experimental Psychology 1961;61:389-399.

5. Cobra D. Image histogram modification based on a new model of visual sustem nonlinearity. J Elect Imaging 1998;7:807-815.

6. Pizer SM, Amburn EP, Austin JD, et al. Adaptive histogram equalization and its variations. Comput Vision Graphics Image Process 1987;39:355-368.

7. Liu H, Nodine CF. A generalized image contrast enhancement technique based on the Heinemann contrast discrimination model. Journal of Electronic Imaging 1996;5:388-395.

8. Mokrane A. A new image contrast enhancement technique based on a contrast discrimination model. Comput Vision Graph Image Process 1992;54:171-180.

LEGENDS

Figure 1. The digital mammography workstation.

Figure 2. The bottom curve is a final MDC lookup table and the top curve is the CRT input-output transfer characteristic. Both curves share a common pixel driving level axis. The non-linearity of the MDC curve is exaggerated for illustrative purposes. The actual difference from the linear curve is usually more subtle. The effect of the MDC lookup table on the displayed image can be seen by following the dotted lines that extrapolate from the image pixel value to the display luminance.

Figure 3. The minimal detectable contrast (MDC) test pattern with typical observer responses indicated by the stars.

Figure 4. The solid lines are examples of two contrast-sensitivity curve from the work of Heinemann (4), one with an adapting luminance at 10 cd/m^2 and the other with an adapting luminance at 100 cd/m^2 . In reality, there is a whole family of curves of similar shape with a minimum at the adapting luminance. The dashed lines are the contrast sensitivities predicted by the parabolic model.

Figure 5. The pattern used for sampling pixel intensities on the breast images. The intensities of the breast are sampled and non-tissue regions beyond the breast are eliminated.

Figure 6. A mammogram image displayed using the standard perceptually linearized display and the MDC tempered display. The arrow shows the skin line.

TABLES

Table 1. A comparison of the area under the ROC curve for three readers who were tested on a set of 75 difficult mammograms using the perceptually linear display and the MDC tempered display.

	Linear Display	Tempered Display	Difference
Reader 1	.910	.930	.020
Reader 2	.861	.869	.008
Reader 3	.627	.750	.123
Mean (sd)	.799	.850	.050 (.063)

Table 2. A comparison of the time to the first decision in seconds for three readers who were tested on a set of 75 difficult mammograms using the perceptually linear display and the MDC tempered display.

	Linear Display	Tempered Display	Difference
Reader 1	76	51	-25
Reader 2	55	84	29
Reader 3	51	47	-4
Mean (sd)	61	61	0 (27)

Figure 1

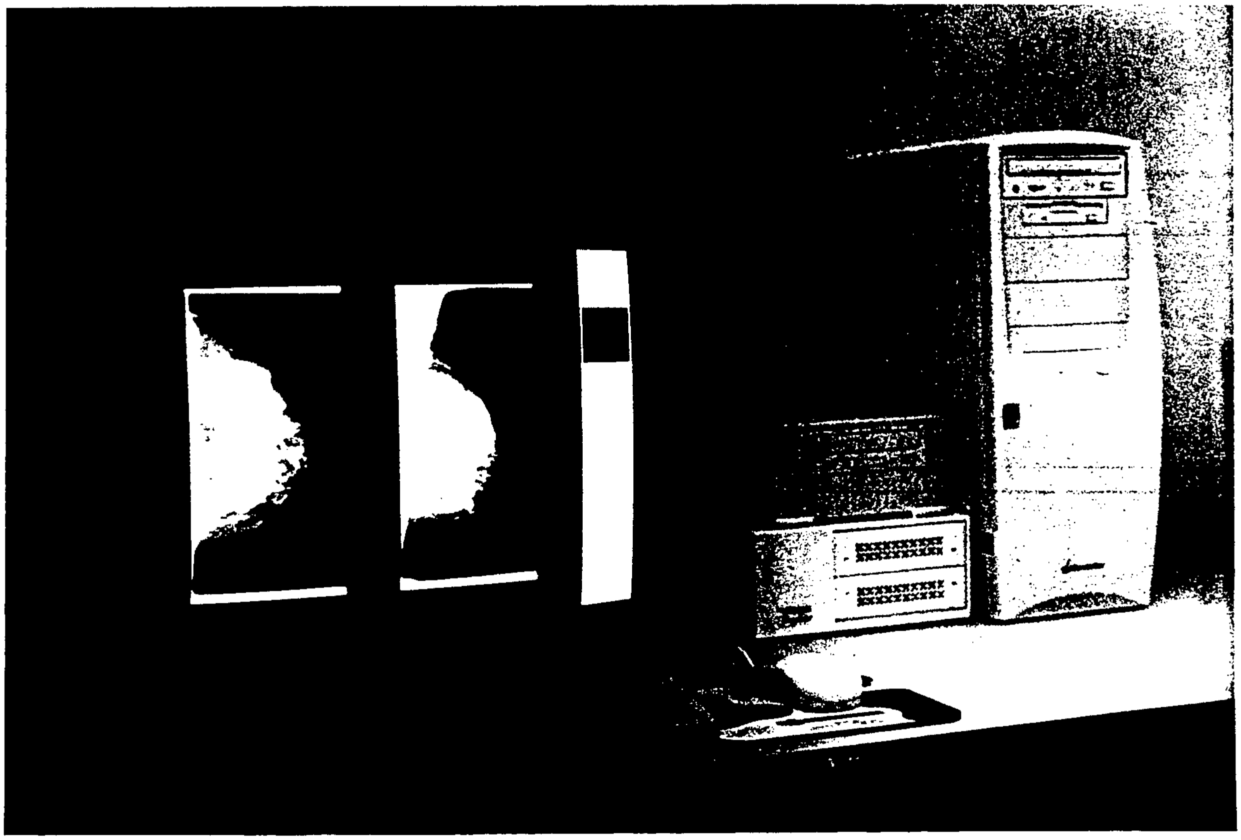


Figure 1. The digital mammography workstation.

Figure 2

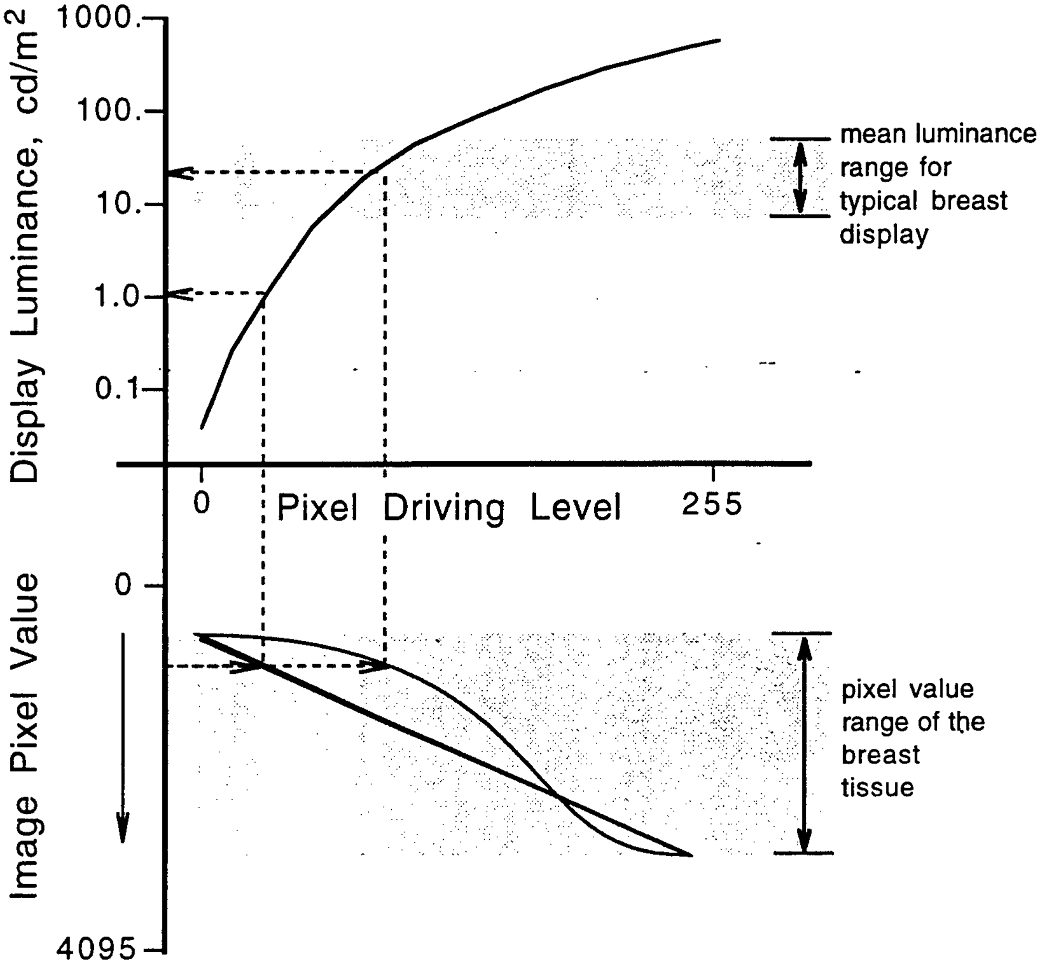


Figure 2. The bottom curve is a final MDC lookup table and the top curve is the CRT input-output transfer characteristic. Both curves share a common pixel driving level axis. The non-linearity of the MDC curve is exaggerated for illustrative purposes. The actual difference from the linear curve is usually more subtle. The effect of the MDC lookup table on the displayed image can be seen by following the dotted lines that extrapolate from the image pixel value to the display luminance.

Figure 3

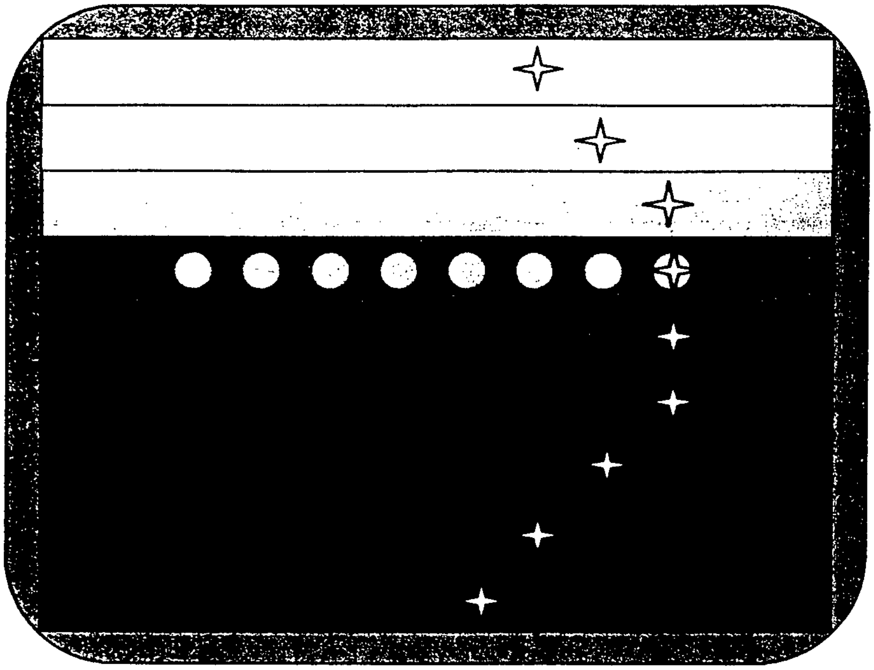


Figure 3. The minimal detectable contrast (MDC) test pattern with typical observer responses indicated by the stars.

Figure 4

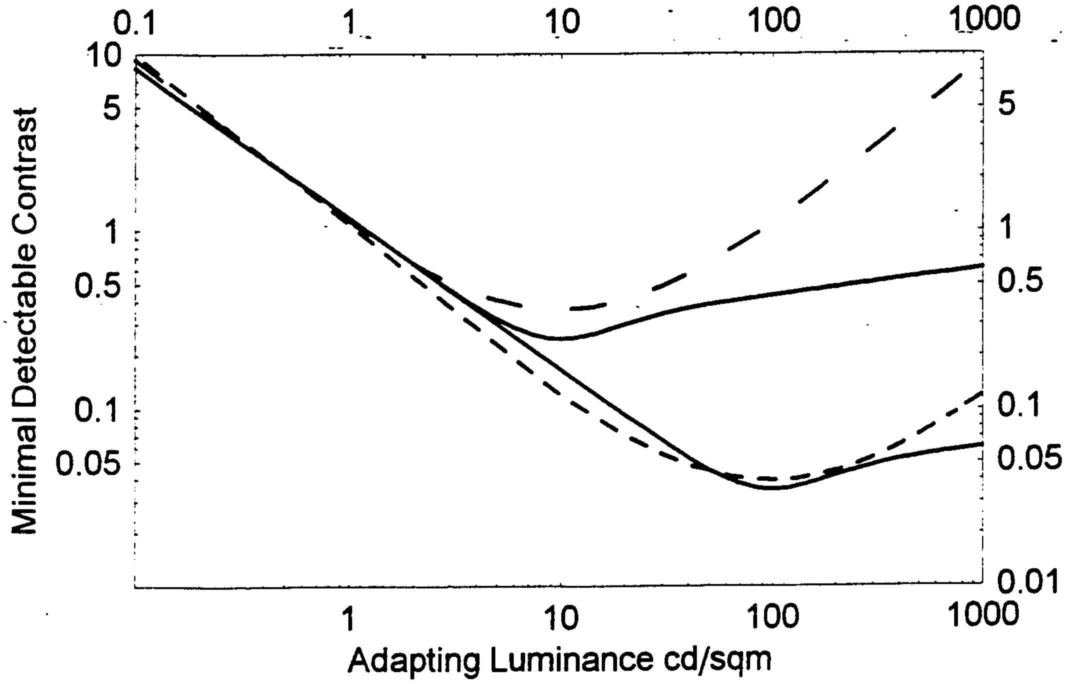


Figure 4. The solid lines are examples of two contrast-sensitivity curve from the work of Heinemann (4), one with an adapting luminance at 10 cd/m^2 and the other with an adapting luminance at 100 cd/m^2 . In reality, there is a whole family of curves of similar shape with a minimum at the adapting luminance. The dotted lines are the contrast sensitivities predicted by the parabolic model.

Figure 5

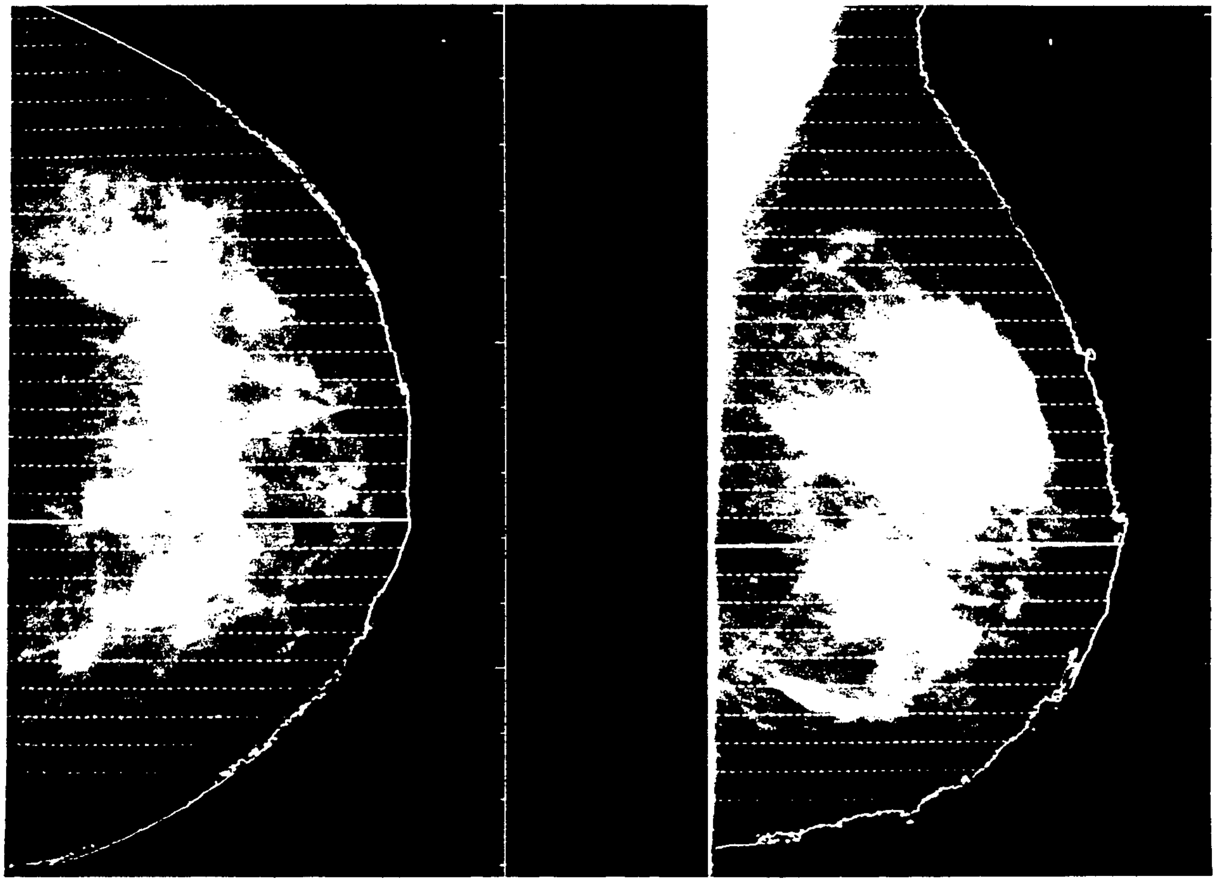


Figure 5. The pattern used for sampling pixel intensities on the breast images. The intensities of the breast are sampled and non-tissue regions beyond the breast are eliminated.

Figure 6

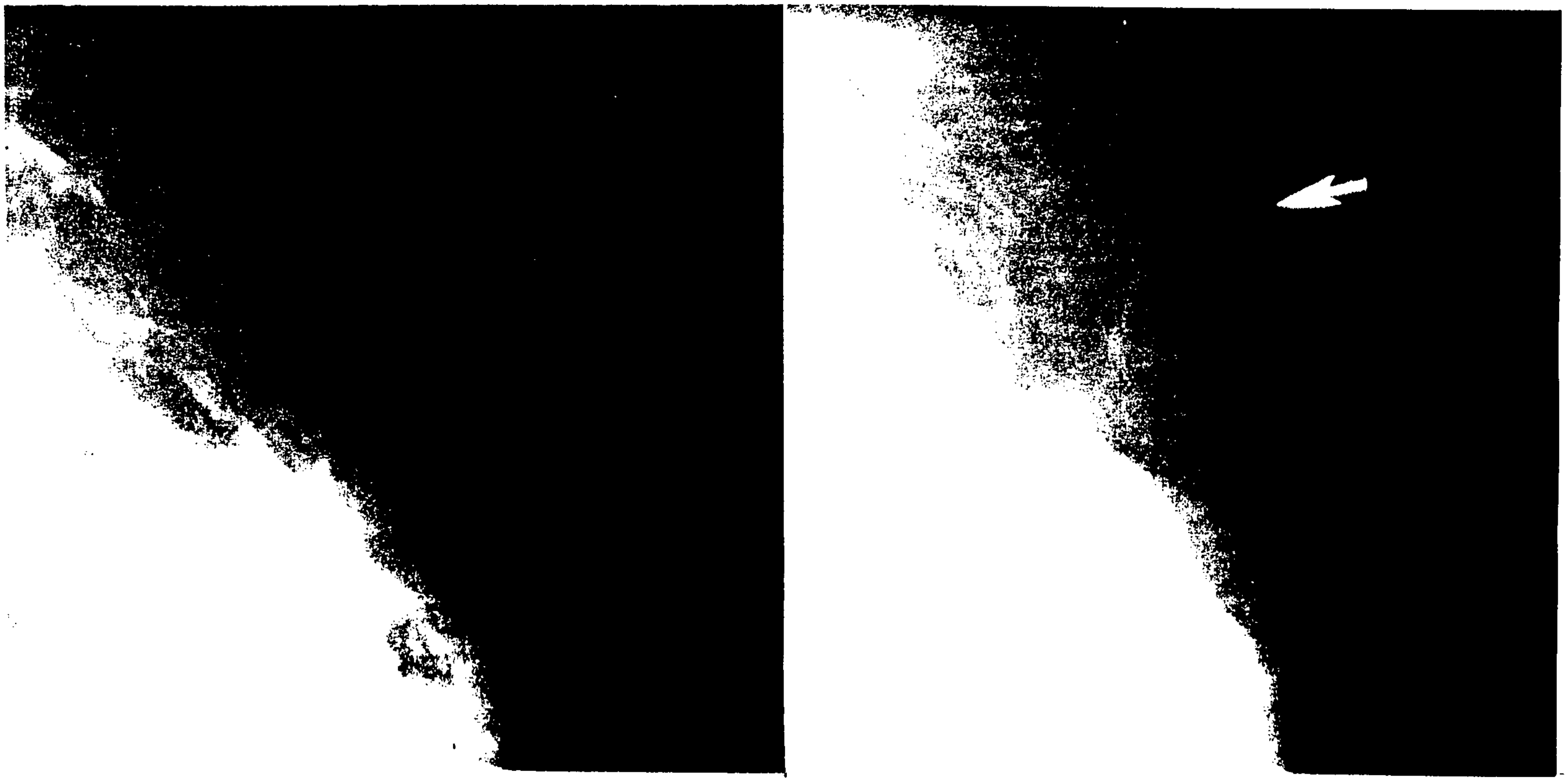


Figure 6. A mammogram image displayed using the standard perceptually linearized display and the MDC tempered display. The arrow shows the skin line.